Lori Cristina Grandin

Aplicações de Modelos Logísticos Regressivos em Biologia Molecular

Dissertação apresentada junto ao Departamento de Estatística do Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas, para obtenção do Título de Mestre em Estatística.

Orientadora:
Profa. Dra. Hildete Prisco Pinheiro

Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Científica
Departamento de Estatística

Campinas - SP

2006

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Lori Cristina Grandin e aprovada pela comissão julgadora.

Campinas, 10 de março de 2006.

Profa. Dra. Hildete Prisco Pinheiro
Departamento de Estatística - UNICAMP
Orientador

Banca Examinadora:

1. Profa. Hildete Prisco Pinheiro (orientadora) - IMECC/UNICAMP
2. Prof. Dr. Luis Aparecido Milan - UFSCar
3. Prof. Dra. Gisela Tunes - IME/USP
4. Prof. Dr. Filidor Edilfonso Vilca Labra (suplente) - IMECC/UNICAMP

Dissertação apresentada junto ao Departamento de Estatística do Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas, para obtenção do Título de Mestre em Estatística.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP
Bibliotecária: Maria Júlia Milani Rodrigues – CRB8a / 2116

Grandin, Lori Cristina
G764a Aplicações de modelos logísticos regressivos em biologia molecular / Lori Cristina Grandin -- Campinas, [S.P. :s.n.], 2006.

Orientadora : Hildete Prisco Pinheiro
Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Seqüência de nucleotídeos. 2. Análise de regressão logística. 3. Análise de seqüência de DNA. I. Pinheiro, Hildete Prisco. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Título em inglês: Logistic regression models with applications in molecular biology

Palavras-chave em inglês (Keywords): 1. Nucleotydes sequence. 2. Logistic regression analysis. 3. DNA sequences analysis.

Área de concentração: Bioestatística

Titulação: Mestre em Estatística

Banca examinadora: Profa. Dra. Hildete Prisco Pinheiro (IMECC-UNICAMP)
Profa. Dra. Gisela Tunes (IME-USP)
Prof. Dr. Luiz Aparecido Milan (UFSCar)

Data da defesa: 10/03/2006

A meus pais, Wilson e Iza.

Agradecimentos

Aos meus pais, Wilson e Iza, por fornecerem todas as condições para que eu me dedicasse aos estudos com tranqüilidade.

À minha orientadora Profa. Hildete Prisco Pinheiro, pelo incentivo, dedicação e amizade desde o desenvolvimento dos projetos da graduação.

Ao Prof. Sérgio Furtado dos Reis, pelo incentivo da aplicação da estatística nos estudos de genética, pelos esclarecimentos na teoria de biologia, pelo carinho e paciência.

Ao Prof. Aluísio Pinheiro, pelas contribuições nos seminários e ao longo do desenvolvimento da dissertação.

Aos professores Gisela Tunes e Luiz Aparecido Milan, por aceitarem participar da banca examinadora, pelas correções e sugestões.

Aos colegas do IMECC, pela amizade e incentivo. Agradeço especialmente à Samara, pela ajuda com o Latex, à Juliana, pelo nascimento de uma amizade que pretendo manter por toda a vida e à Camila e Jaqueline, pelos momentos de descontração e estudo conjunto.

Às amizades que fiz durante o período do meu mestrado e que desejo manter por toda a vida.

Ao CNPQ, pelo suporte financeiro, fundamental para o desenvolvimento desse projeto.

"Somos feitos do tecido de que são feitos os sonhos"

Shakespeare

Resumo

O avanço do sequenciamento dos genes tem incentivado o desenvolvimento de novas técnicas estatísticas para analisar dados genéticos. Nesse trabalho, os modelos logísticos regressivos, introduzidos por Bonney (1986), são apresentados primeiramente no contexto de análise de dados de família e posteriormente esses modelos são utilizados para analisar freqüências de códons em seqüências de DNA mitocondrial.

Considerar independência entre os nucleotídeos no códon pode ser uma suposição muito forte, ou seja, biologicamente irreal. Por isso, várias estruturas de dependência são apresentadas para analisar as freqüências dos códons. Por exemplo, uma estrutura markoviana de primeira ordem pode ser adequada para explicar a dependência das bases no códon. A função de log-verossimilhança é avaliada e várias comparações são feitas para analisar qual o modelo mais parcimonioso. Aplicações desses modelos são feitas utilizando-se dados reais de seqüências do gene NADH4 do genoma mitocondrial humano.

Abstract

The advance of gene sequencing has stimulated the development of new statistical techniques to analyze genetic data. In this work the logistic regressive models, introduced by Bonney (1986), are presented first in the context of analysis of familial data and then they are used to analyze codon frequencies in mitochondrial DNA sequences.

The assumption of independence among nucleotide frequencies in a codon can be a very strong one, or biologically unreal. In view of this, several structures of dependence are presented to analyze the codon frequencies. For example, a first order Markovian structure can be appropriate to explain the dependence of the base frequencies in the codon. The log-likelihood function is evaluated and several comparisons are made to analyze which is the most parcimonious model. Applications of these models are made using real data of NADH4 gene sequences of the human mitochondrial genome.

Conteúdo

Lista de Tabelas p. xvii

Lista de Figuras p. xix

1 Introdução e Revisão de Conceitos Biológicos p. 1

1.1 Motivação . . . . . p. 1

1.2 Conceitos Básicos de Biologia Molecular . . . . . p. 2

1.2.1 Ácidos Nucléicos . . . . . p. 2

1.2.2 Ácido Desoxirribonucléico (DNA) . . . . . p. 3

1.2.3 Ácido Ribonucléico (RNA) . . . . . p. 4

1.2.4 O Código Genético . . . . . p. 5

1.2.5 Síntese de Proteínas . . . . . p. 8

1.2.6 Dogma Central . . . . . p. 9

1.2.7 Mitocôndrias . . . . . p. 10

1.2.8 Mutação . . . . . p. 13

1.2.9 Substituição de Nucleotídeos . . . . . p. 13

2 Modelos Logísticos Regressivos para Dados de Família p. 17

2.1 Introdução . . . . . p. 17

2.2 Modelo de Regressão Logística . . . . . p. 18
2.2.1 Ajuste do Modelo Logístico por Máxima Verossimilhança . p. 19
2.3 Modelos Logísticos Regressivos . . . . . p. 22
2.3.1 Modelos para Observações Binárias Dependentes . . . . . p. 25
2.3.2 Cálculo da Função de Verossimilhança . . . . . p. 34

3 Modelos para Analisar Seqüências de DNA p. 37
3.1 Introdução . . . . . p. 37
3.2 Modelos Logísticos Regressivos para Dados Politômicos . . . . . p. 39
3.3 Verossimilhanças e Matriz de Informação . . . . . p. 44

4 Aplicação p. 63
4.1 Neuropatia Ótica Hereditária de Leber . . . . . p. 64
4.1.1 Comparações entre a SRC e Demais Seqüências . . . . . p. 65
4.2 Ajuste de Modelos para a Seqüência de Referência de Cambridge . p. 68
4.2.1 Teste da Razão de Verossimilhanças para o Modelo Aditivo p. 69
4.2.2 Comparação entre os Modelos . . . . . p. 72
4.3 Ajuste dos Modelos para várias Seqüências de DNA . . . . . p. 75
4.3.1 Diagnóstico dos Modelos . . . . . p. 80

5 Considerações Finais p. 85

Referências p. 89

Apêndice p. 93

Lista de Tabelas

1.1 Aminoácidos . . . . . p. 5

1.2 Código Genético Universal - Códon e aminoácido correspondente p. 6

1.3 Código Genético Mitocondrial para Mamíferos - Códon e aminoácido correspondente . . . . . p. 7

1.4 Freqüências Relativas dos diferentes tipos de substituições mutacionais em uma seqüência aleatória que codifica proteínas. . . . . p. 15

3.5 Codificação . . . . . p. 41

4.6 Descrição das Mutações Encontradas comparando-se a SRC com as demais seqüências . . . . . p. 67

4.7 Teste da Razão de Verossimilhanças (RV) . . . . . p. 71

4.8 Ajuste dos Modelos para o Gene NADH4, SRC . . . . . p. 72

4.9 Estimativa dos Parâmetros do Modelo de Markov com Covariáveis (3) . . . . . p. 74

4.10 Frequências Relativas das bases por posição no códon para a SRC p. 76

4.11 Ajuste dos Modelos para o Gene NADH4, n=30 seqüências . . . . p. 77

4.12 Ajuste do Modelo Aditivo para várias combinações de covariáveis, n=30 seqüências . . . . . p. 79

4.13 Estimativa dos parâmetros do Modelo Aditivo com Covariáveis (1) p. 80

4.14 SQR para os modelos ajustados para n= 30 seqüências . . . . . p. 81

A.15 Número e Posição de mutações encontradas comparando a SRC com indivíduos doentes e não-doentes. . . . . p. 97

A.16 Continuação. . . . . p. 98

A.17 Frequências Observadas da SRC do gene NADH4 e Valores das Covariáveis . . . . . p. 99

A.18 Probabilidades Observadas e Estimadas para os modelos com Covariáveis para n= 30 seqüências . . . . . p. 100

A.19 Continuação . . . . . p. 101

A.20 Probabilidades Observadas e Estimadas para os modelos sem Covariáveis para n= 30 seqüências . . . . . p. 102

A.21 Continuação . . . . . p. 103

Lista de Figuras

1.1 Dupla hélice do DNA. . . . . p. 3

1.2 Estrutura de uma molécula de DNA. . . . . p. 3

1.3 Processo de expressão gênica: Síntese de Proteínas. . . . . p. 8

1.4 Dogma Central. . . . . p. 9

1.5 Estrutura da Mitocôndria. . . . . p. 11

1.6 Genes do DNA Mitocondrial. . . . . p. 12

4.7 Modelo Aditivo com e sem Covariáveis, n = 30 seqs. . . . . p. 82

4.8 Modelo de Markov com e sem Covariáveis, n = 30 seqs. . . . . p. 82

4.9 Igualmente Preditvo com e sem Covariáveis, n = 30 seqs. . . . . p. 83

4.10 Independente com e sem Covariáveis, n = 30 seqs. . . . . p. 83

1 Introdução e Revisão de Conceitos Biológicos

1.1 Motivação

As técnicas moleculares e a tecnologia computacional têm avançado muito ultimamente, o que tem facilitado o sequenciamento de uma quantidade muito grande de genes, além do armazenamento dessas informações em grandes bancos de dados no mundo inteiro. Com todo esse avanço e com o crescente interesse dos pesquisadores no sequenciamento do DNA, as análises estatísticas têm tido grande importância e interesse.

Com o início do Projeto Genoma Humano em 1990 e subseqüente disponibilização de seqüenciadores automáticos de DNA capazes de gerar dados genômicos em grande escala, surge a necessidade de novas ferramentas estatísticas para a análise desses bancos de dados.

Neste capítulo apresentaremos alguns conceitos básicos de Biologia Molecular para o melhor entendimento do assunto abordado nessa dissertação. No capítulo 2, será feita uma breve revisão de modelos de regressão logística e serão apresentados os modelos logísticos regressivos para analisar dados de família. No capítulo 3, apresentaremos os modelos logísticos regressivos para analisar freqüências de códons em seqüências de DNA, ou seja, serão utilizadas covariáveis e estrutura de dependência entre os nucleotídeos para explicar as freqüências de códons em seqüências de DNA mitocondrial. A função de log-verossimilhança e a matriz de

informação de Fisher para o modelo logístico regressivo politômico nominal serão apresentadas. As desvantagens desses modelos serão abordadas, bem como alternativas à eles. Aplicações dos modelos logísticos regressivos em biologia molecular serão mostradas no capítulo 4 utilizando-se a seqüência de referência de Cambridge e várias outras seqüências de DNA mitocondrial humano. Faremos uma comparação entre oito modelos ajustados à dados reais, utilizando o critério de Akaike e o teste da Razão de Verossimilhanças. Algumas medidas alternativas de diagnóstico serão aplicadas para averiguar qual modelo se ajusta melhor aos dados. No capítulo 5, algumas conclusões sobre a utilização dos modelos logísticos regressivos aplicados em biologia molecular são apresentadas e discutiremos sobre abordagens alternativas à esses modelos.

1.2 Conceitos Básicos de Biologia Molecular

A Biologia Molecular retrata o estudo das células e moléculas, blocos básicos utilizados na construção de todas as formas de vida (Casley, 1992). Em particular, estuda-se o genoma dos organismos, definido como o conjunto de suas informações genéticas.

1.2.1 Ácidos Nucléicos

Os ácidos nucléicos são macromoléculas de extrema importância biológica em todos os organismos vivos. Eles dão instruções sobre quais proteínas serão sintetizadas e em que quantidade, e também estocam e transmitem a informação genética da célula. Existem dois tipos de ácidos nucléicos: o **DNA** (ácido desoxirribonucléico) e o **RNA** (ácido ribonucléico).

1.2.2 Ácido Desoxirribonucléico (DNA)

O **DNA** é chamado de "molécula da vida"pois contém o código para a construção das proteínas em todos os seres vivos. Seres eucariontes são aqueles cujas células são eucarióticas, isto é, as células possuem núcleo delimitado por uma membrana, a carioteca. Nos eucariontes, o DNA é encontrado no núcleo celular formando os cromossomos e também nas mitocôndrias e nos cloroplastos.

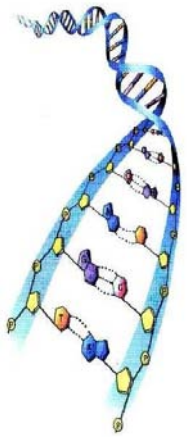

Figura 1.1: Dupla hélice do DNA.

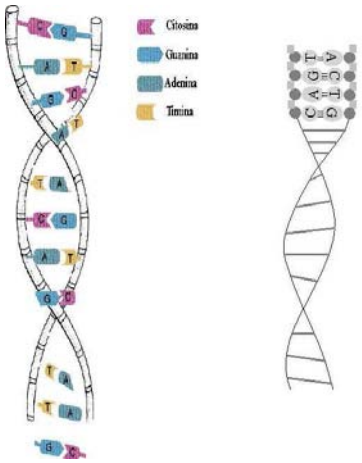

Figura 1.2: Estrutura de uma molécula de DNA.

Uma molécula de DNA consiste de duas fitas anti-paralelas entrelaçadas em forma de dupla hélice, conforme pode ser visualizado nas Figuras 1.1 e 1.2. Este modelo da "dupla hélice enrolada"foi estabelecido em 1953 por James Watson e Francis Crick.

Nucleotídeos são as unidades fundamentais dos genomas e cada nucleotídeo consiste de um açúcar, um fosfato e uma base (Andrade & Pinheiro, 2002). Cada fita é composta por uma seqüência de nucleotídeos (bases), que podem ser de quatro tipos: **Adenina (A)**, **Guanina (G)**, **Citosina (C)** e **Timina (T)**. Cada nucleotídeo de uma fita se liga a outro complementar da segunda, de tal modo que Adenina pareia-se com a Timina e a Citosina pareia-se com a Guanina. Como resultado dessas regras de pareamento, as duas seqüências de bases que formam as moléculas de DNA são complementares entre si.

Os pares de bases (pb) são a unidade de tamanho do DNA. O comprimento de uma molécula de DNA normalmente é muito grande, como por exemplo a célula humana cuja as moléculas de DNA tem centenas de milhões de pb.

1.2.3 Ácido Ribonucléico (RNA)

O **RNA** (ácido ribonucléico) é uma molécula formada geralmente por uma única fita. Sua estrutura é similar ao DNA, exceto pelo fato de que no lugar do açúcar desoxirribose, o RNA contém o açúcar ribose, no lugar da base Timina, o RNA tem a base Uracila. Existem três classes de RNA:

- **RNA mensageiro (mRNA)**: contém a informação genética para a seqüência de aminoácidos. O RNA mensageiro é formado no núcleo e contém a "mensagem", o código transcrito a partir do DNA, para a síntese das proteínas.
- **RNA transportador (tRNA)**: identifica e transporta as moléculas de aminoácido até o ribossomo, ou seja, o tRNA está presente no citoplasma e é responsável pelo transporte dos aminoácidos até os ribossomos para a síntese protéica.

- **RNA ribossômico (rRNA)**: facilita a interação das outras moléculas de RNA na síntese protéica, ou seja, o rRNA faz parte da estrutura dos ribossomos e participa do processo de tradução dos códons para construção das proteínas.

1.2.4 O Código Genético

Cada grupo adjacente de 3 nucleotídeos constitui um **códon** que especifica um aminoácido correspondente na cadeia polipeptídica (ou protéica). O nome códon foi dado justamente para especificar que algumas das triplas de bases codificam aminoácidos.

Tabela 1.1: Aminoácidos

Aminoácidos	Sigla	Códons Sinônimos
Alanina	Ala	GCT, GCC, GCA, GCG
Arginina	Arg	CGT, CGC, CGA, CGG,AGA,AGG
Asparagina	Asn	AAT, AAC
Ácido Aspártico	Asp	GAT, GAC
Cisteina	Cys	TGT, TGC
Ácido Glutâmico	Glu	GAA, GAG
Glutamina	Gln	CAA, CAG
Glicina	Gly	GGT, GGC, GGA, GGG
Histidina	His	TAT, CAC
Isoleucina	Ile	ATT, ATC, ATA
Leucina	Leu	TTA, TTG, CTT, CTC, CTA, CTG
Lisina	Lys	AAA, AAG
Metionina	Met	ATG
Fenilalanina	Phe	TTT, TTC
Prolina	Pro	CCT, CCC, CCA, CCG
Serina	Ser	TCT, TCC, TCA, TCG, AGT, AGC
Treonina	Thr	ACT, ACA, ACG, ACC
Triptofano	Trp	TGG
Tirosina	Tyr	TAT, TAC
Valina	Val	GTT, GTC, GTA, GTG

Como existem 4 nucleotídeos que formam o DNA, então existem 64 possíveis

combinações de triplas de nucleotídeos, ou seja, 64 códons. No entanto, sabe-se que os códons codificam um total de 20 aminoácidos na cadeia polipeptídica, ou seja, códons diferentes podem codificar o mesmo aminoácido. Os códons que codificam o mesmo aminoácido são chamados de **códons sinônimos**. Em geral, os códons sinônimos possuem as duas primeiras posições do códon iguais. Os 20 aminoácidos e suas siglas encontram-se na Tabela 1.1. Para um gene [1] específico, não são sintetizados as 64 possíveis combinações das bases que formam os códons e portanto, algumas combinações não sintetizam nenhum aminoácido. Esse é o caso dos **terminadores** ou códons de parada, ou seja, códons que especificam o término da síntese polipeptídica e não sintetizam nenhum aminoácido. Esses códons significam simplesmente o final da seqüência. Além dos códons de parada, existe também o códon que indica o início da síntese polipeptídica, que é em geral o códon que corresponde ao aminoácido **Metionina**. A correspondência entre cada trio de nucleotídeos e seu correspondente aminoácido é conhecido como **código genético**.

Tabela 1.2: Código Genético Universal - Códon e aminoácido correspondente

TTT	Phe/F	**TCT**	Ser/S	**TAT**	Tyr/Y	**TGT**	Cys/C
TTC	Phe/F	**TCC**	Ser/S	**TAC**	Tyr/Y	**TGC**	Cys/C
TTA	Leu/L	**TCA**	Ser/S	**TAA**	*Ter*	**TGA**	*Ter*
TTG	Leu/L	**TCG**	Ser/S	**TAG**	*Ter*	**TGG**	Trp/W
CTT	Leu/L	**CCT**	Pro/P	**CAT**	His/H	**CGT**	Arg/R
CTC	Leu/L	**CCC**	Pro/P	**CAC**	His/H	**CGC**	Arg/R
CTA	Leu/L	**CCA**	Pro/P	**CAA**	Gln/Q	**CGA**	Arg/R
CTG	Leu/L	**CCG**	Pro/P	**CAG**	Gln/Q	**CGG**	Arg/R
ATT	Ile/I	**ACT**	Thr/T	**AAT**	Asn/N	**AGT**	Ser/S
ATC	Ile/I	**ACC**	Thr/T	**AAC**	Asn/N	**AGC**	Ser/S
ATA	Ile/I	**ACA**	Thr/T	**AAA**	Lys/K	**AGA**	Arg/R
ATG	*Met/M*	**ACG**	Thr/T	**AAG**	Lys/K	**AGG**	Arg/R
GTT	Val/V	**GCT**	Ala/A	**GAT**	Asp/D	**GGT**	Gly/G
GTC	Val/V	**GCC**	Ala/A	**GAC**	Asp/D	**GGC**	Gly/G
GTA	Val/V	**GCA**	Ala/A	**GAA**	Glu/E	**GGA**	Gly/G
GTG	Val/V	**GCG**	Ala/A	**GAG**	Glu/E	**GGG**	Gly/G

[1] Um gene é um segmento de DNA que contém informações para a síntese de uma ou mais proteínas.

As diferentes codificações podem ser visualizadas na Tabela 1.2. Esse código genético é usado na maioria dos seres vivos, daí o nome código genético universal.

Com poucas exceções, o código genético para genes que codificam proteínas é universal, ou seja, são utilizadas as mesmas regras para determinação de genes eucarióticos. No entanto, o genoma mitocondrial usa um código que é diferente do código genético universal. Porém, essas diferenças são pequenas. O código genético mitocondrial para mamíferos pode ser visto na Tabela 1.3. O fato de ser possível traduzir genes de um organismo em outro, p. ex., genes humanos, em E. coli, sugeria que o código padrão apresentado na Tabela 1.2 era universal. Todavia, o estudo de diferentes seqüências de DNA a partir dos anos 80 revelaram algumas divergências em relação ao padrão.

Tabela 1.3: Código Genético Mitocondrial para Mamíferos - Códon e aminoácido correspondente

TTT	Phe/F	**TCT**	Ser/S	**TAT**	Tyr/Y	**TGT**	Cys/C
TTC	Phe/F	**TCC**	Ser/S	**TAC**	Tyr/Y	**TGC**	Cys/C
TTA	Leu/L	**TCA**	Ser/S	**TAA**	Ter/T	**TGA**	**Trp/W**
TTG	Leu/L	**TCG**	Ser/S	**TAG**	Ter/T	**TGG**	Trp/W
CTT	Leu/L	**CCT**	Pro/P	**CAT**	His/H	**CGT**	Arg/R
CTC	Leu/L	**CCC**	Pro/P	**CAC**	His/H	**CGC**	Arg/R
CTA	Leu/L	**CCA**	Pro/P	**CAA**	Gln/Q	**CGA**	Arg/R
CTG	Leu/L	**CCG**	Pro/P	**CAG**	Gln/Q	**CGG**	Arg/R
ATT	Ile/I	**ACT**	Thr/T	**AAT**	Asn/N	**AGT**	Ser/S
ATC	Ile/I	**ACC**	Thr/T	**AAC**	Asn/N	**AGC**	Ser/S
ATA	**Met/M**	**ACA**	Thr/T	**AAA**	Lys/K	**AGA**	**Ter/T**
ATG	Met/M	**ACG**	Thr/T	**AAG**	Lys/K	**AGG**	**Ter/T**
GTT	Val/V	**GCT**	Ala/A	**GAT**	Asp/D	**GGT**	Gly/G
GTC	Val/V	**GCC**	Ala/A	**GAC**	Asp/D	**GGC**	Gly/G
GTA	Val/V	**GCA**	Ala/A	**GAA**	Glu/E	**GGA**	Gly/G
GTG	Val/V	**GCG**	Ala/A	**GAG**	Glu/E	**GGG**	Gly/G

As diferenças do código genético mitocondrial de mamíferos e do código genético universal encontram-se em negrito na Tabela 1.3. Por exemplo, em mitocôndrias

de mamíferos o códon para a Metionina (Met) iniciadora pode ser AUG ou AUA (Ile no padrão); UGA especifica Trp e não terminação; AGA e AGG especificam terminação e não Arg.

1.2.5 Síntese de Proteínas

Um fragmento de DNA pode conter diversos genes. A propriedade mais importante dos genes está no fato de que eles codificam proteínas, componentes essenciais de todo ser vivo. As proteínas possuem diversas funções biológicas [Lewis, 2001]. Elas podem ter papel estrutural, como no caso do colágeno presente nos tendões, ou estar ligadas a atividades regulatórias, como no caso das enzimas, que catalisam diversas reações químicas nas células. As proteínas também são seqüências lineares, compostas de conjuntos de aminoácidos. As **proteínas** são cadeias de aminoácidos, e portanto, para especificar uma proteína, deve-se especificar quais aminoácidos ela contém. O processo pelo qual as seqüências de nucleotídeos dos genes são interpretadas na produção de proteínas é denominado **expressão gênica** (Figura 1.3).

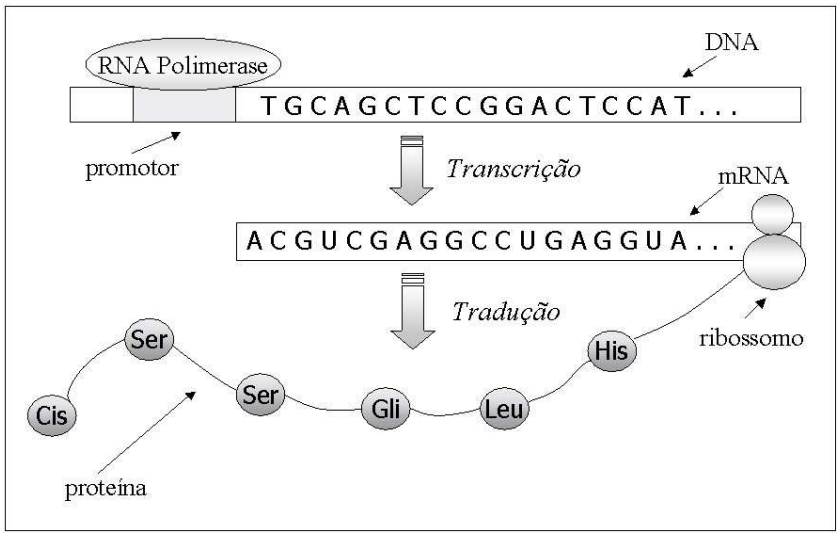

Figura 1.3: Processo de expressão gênica: Síntese de Proteínas.

A expressão gênica é composta por duas etapas: na primeira, denominada transcrição, uma enzima de RNA polimerase se liga a uma região do DNA denominada promotora e inicia a síntese de um RNA mensageiro (mRNA). Na segunda etapa da expressão, chamada tradução, é realizada a síntese da molécula de proteína, a partir do mRNA. Cada grupo de três nucleotídeos do mRNA representa um aminoácido, constituinte de uma proteína.

1.2.6 Dogma Central

O Dogma Central foi estabelecido em 1956 por Francis Crick na tentativa de relacionar o DNA, o RNA e as proteínas. Francis postulava que o sentido da construção das moléculas é sempre de DNA à Proteína (fluxo unidirecional da informação). O DNA pode se replicar e dar origem a novas moléculas de DNA, pode ainda ser transcrito em RNA, e este por sua vez traduz o código genético em proteínas.

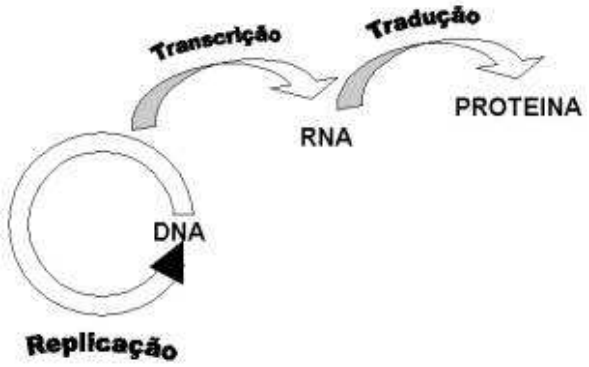

Figura 1.4: Dogma Central.

O ciclo do DNA para ele mesmo significa que a molécula pode ser copiada. Isto é chamado replicação. O passo do DNA ao RNA é chamado transcrição e a formação de proteína é chamada tradução.

Em organismos eucariotos, algumas partes da molécula de mRNA não são traduzidas em proteínas. O material genético dos organismos eucariotos possui, portanto, seqüências de nucleotídeos que são codificadas em proteínas, os **exons**, e

seqüências que não participam desse processo, os **introns**. As fronteiras entre essas seqüências são denominadas sítios de **splicing**, nome decorrente do processamento no qual os introns são removidos da molécula de mRNA.

Somente 1% do genoma humano consiste de regiões codificadoras. O genoma humano tem entre 30.000 e 40.000 genes. Os exons compreendem aproximadamente 5% do gene. A estimativa mais recente do número total de genes codificadores de proteínas existentes no genoma humano é de 20.000 a 25.000, bastante inferior ao número inicialmente previsto de mais de 100.000. Os introns não são encontrados nos genes mitocondriais de mamíferos, mas são a regra entre os genes nucleares.

1.2.7 Mitocôndrias

As **mitocôndrias** (do grego mito: filamento e chondrion: grânulo) estão presentes no citoplasma das células eucarióticas, sendo caracterizadas por uma série de propriedades morfológicas, bioquímicas e funcionais. Geralmente, são estruturas cilíndricas com aproximadamente 0,5 micrômetros de diâmetro e vários micrômetros de comprimento. As mitocôndrias são organelas notavelmente móveis e plásticas, mudando constantemente suas formas e mesmo fundindo-se umas com as outras e separando-se novamente. Possuem organização estrutural e composição lipoprotéica características, e contêm um grande número de enzimas e coenzimas que participam das reações de transformação da energia celular. A mitocôndria tem como função realizar a maior parte das oxidações celulares e produzir a massa de ATP (Adenosina Trifosfato), ou seja, a energia celular das células animais.

A maior parte do DNA é constituinte dos cromossomos, localizados nos núcleos das células, e o restante se encontra em estruturas denominadas de mitocôndrias, situadas no citoplasma celular.

Células de eucariontes apresentam DNA nuclear e DNA de organelas, tais como mitocôndrias em animais e cloroplastos em plantas denominados "genomas extranucleares". No reino animal, a mitocôndria é a única organela que contém seu próprio

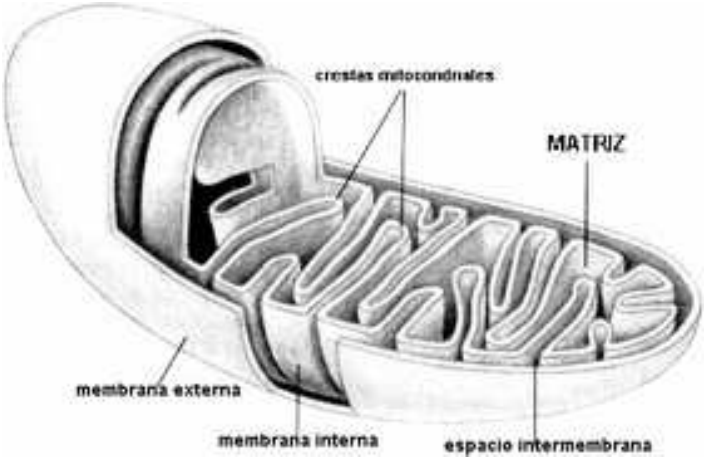

Figura 1.5: Estrutura da Mitocôndria.

DNA, com genes que codificam várias moléculas integrantes da cadeia respiratória e genes que codificam RNA ribossomal e RNA transportador (tRNA).

Em termos genéticos ou fisiológicos, a mitocôndria é uma organela semi-autônoma. Ela tem seu próprio genoma, que no homem apresenta 37 genes:

- dois genes codificam RNAs ribossomais;
- 22 codificam RNAs transportadores;
- 13 codificam polipeptídeos que integram alguns componentes da cadeia respiratória e da ATP-sintetase, ou seja, esses genes codificam RNAs mensageiros para proteínas envolvidas diretamente no processo de transporte de elétrons e fosforilação oxidativa. São 13 seqüências codificantes de proteínas. Dessas 13, têm-se:

  3 são subunidades do complexo **citocromo oxidase**

  2 são subunidades da **ATPase**

  7 são subunidades do complexo **NADH** (CoQ reductase)

  1 (cit b) é a subunidade do complexo CoQ, **citocromo reductase**

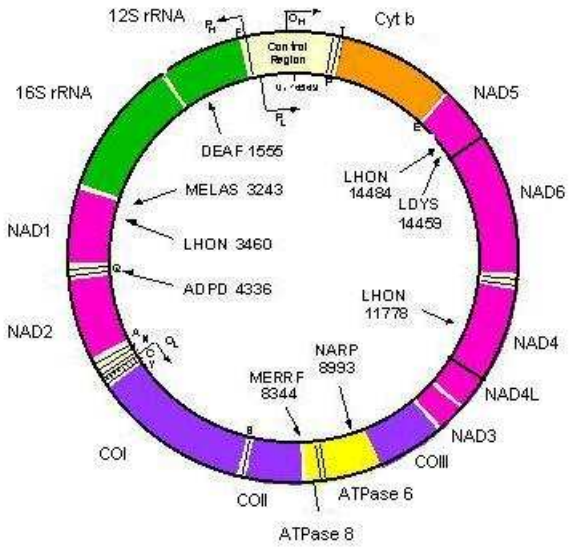

Figura 1.6: Genes do DNA Mitocondrial.

O DNA mitocondrial (DNAmt) humano, localizado na matriz da mitocôndria, contém 16.569 pares de bases, um dos menores que se conhece. O DNAmt humano é um pequeno DNA circular presente dentro das mitocôndrias (as usinas energéticas da célula) no citoplasma. Este DNA tem uma série de características genéticas peculiares, destacando-se o fato de ter herança puramente materna. Em outras palavras, todo o DNA mitocondrial de um indivíduo vem de sua mãe apenas, sem nenhuma contribuição paterna. A mãe transmite o DNA de suas mitocôndrias a todos os seus filhos. Suas filhas, por sua vez, o transmitem, mas seus filhos não.

Os genes de DNAmt mais comumente usados para análises estatísticas são:

- **12S rRNA**: Esta é a menor sub-unidade dos ribossomos mitocondriais. Como a maioria dos genes ribossômicos, este gene é bem conservado, embora certas regiões apresentem alta taxa de substituição.

- **16S rRNA**: Esta é a maior sub-unidade dos ribossomos mitocondriais. Este gene, como o 12S é bem conservado, sua estrutura secundária parece ser mais

conservada ainda. Sua região com alta taxa de substituição permite fazer estudos populacionais.

- **Citocromo oxidade I (Cox I)**: Este gene codifica uma proteína que faz parte da cadeia de transporte de elétrons na mitocôndria. Este gene é bem conservado, muitas de suas sustituições ocorrem sem mudanças de aminoácidos.

- **Citocromo b**: Como o Cox I, a proteína codificada por este gene faz parte da cadeia de transporte de elétrons na mitocôndria. Este talvez seja o gene mais comumente usado em inferências filogenéticas. Apresenta grande variabilidade de taxas de evolução possibilitando inferências em diversos níveis.

1.2.8 Mutação

As seqüências de DNA são normalmente copiadas exatamente durante o processo de replicação cromossômica. No entanto, raramente ocorrem erros que geram seqüências diferentes da original. Esses erros são chamados de **mutações**. As mutações podem ser classificadas de acordo com o comprimento da seqüência de DNA afetada pela mutação. Por exemplo, as mutações podem afetar um único nucleotídeo (mutação pontual) ou vários nucleotídeos adjacentes. As mutações também podem ser classificadas de acordo com o tipo de mudança causada pela mutação, ou seja, **substituição**, a substituição de um nucleotídeo por outro, **deleção**, a remoção de um ou mais nucleotídeos da seqüência de DNA, **inserção**, a adição de um ou mais nucleotídeos na seqüência. A taxa de mutação do DNAmt é cerca de dez vezes maior que a do DNA nuclear.

1.2.9 Substituição de Nucleotídeos

As substituições de nucleotídeos estão divididas em **transições** e **transversões**. Transições são substituições entre A e G (purinas) ou entre C e T (pirimidinas). Transversões são substituições entre uma purina e uma pirimidina.

As substituições de nucleotídeos que ocorrem nas regiões que codificam proteína podem ser caracterizadas pelo efeito no produto da translação, ou seja, na proteína. A substituição ou a mutação é dita **sinônima** ou silenciosa se ela não causa mudança de aminoácido. Caso contrário, a mutação é dita **não-sinônima**. Mutações não-sinônimas ou mutações que causam mudança de aminoácido são classificadas em mutação **efetiva (missense)** e **não-efetiva (nonsense)**. Uma mutação missense muda o códon afetado para um outro códon que especifica um aminoácido diferente do previamente codificado. Uma mutação nonsense muda o códon para um codon de terminação, assim terminando prematuramente o processo de translação e provocando a produção de uma proteína truncada.

Cada um dos códons efetivos pode mutar com outros nove códons levando em consideração uma única substituição. Por exemplo, para o códon CCT(Pro) podem ocorrer seis substituições não-sinônimas, para TCT(Ser), ACT(Thr), GCT(Ala), CTT(Leu), CAT(His), CGT(Arg), e três mutações sinônimas, para CCC, CCA, CCG. Como o código genético universal tem 61 códons efetivos, existem $61 \times 9 = 549$ possíveis substituições de nucleotídeos. Supondo que as substituições de nucleotídeo acontecem aleatoriamente e que todos os códons são igualmente frequentes em regiões codificadoras, pode-se calcular a proporção esperada dos diferentes tipos de substituição de nucleotídeos do código genético. Para o código genético mitocondrial tem-se 60 códons efetivos, então, tem-se $60 \times 9 = 540$ possíveis substituições de nucleotídeos. A Tabela 1.4 mostra os resultados para ambos códigos genéticos.

Por causa da estrututa do código genético, substituições sinônimas ocorrem principalmente na terceira posição dos códons. Observando a Tabela 1.4 é fácil notar que para o código genético universal, quase 70% de todas as substituições possíveis de nucleotídeos na terceira posição são mutações sinônimas. Por sua vez, todas as mudanças de nucleotídeos na segunda posição dos códons são não-sinônimas. Para a primeira posição tem-se que 96% de todas as mudanças de nucleotídeo possíveis são não-sinônimas.

Tabela 1.4: Freqüências Relativas dos diferentes tipos de substituições mutacionais em uma seqüência aleatória que codifica proteínas.

	Código Universal		Código Mitocondrial	
Substituição	Número	Percentil	Número	Percentil
Total em todos os códons	549	100	540	100
Sinônima	134	25	128	24
Não-sinônima	415	75	412	76
Efetiva	392	71	384	71
Não-Efetiva	23	4	28	5
Total na primeira posição	183	100	180	100
Sinônima	8	4	4	2
Não-sinônima	175	96	176	98
Efetiva	166	91	164	91
Não-Efetiva	9	5	12	7
Total na segunda posição	183	100	180	100
Sinônima	0	0	0	0
Não-sinônima	183	100	180	100
Efetiva	176	96	172	96
Não-Efetiva	7	4	8	4
Total na terceira posição	183	100	180	100
Sinônima	126	69	124	69
Não-sinônima	57	31	56	31
Efetiva	50	27	48	27
Não-Efetiva	7	4	8	4

2 *Modelos Logísticos Regressivos para Dados de Família*

2.1 Introdução

Os modelos logísticos regressivos foram primeiramente introduzidos no contexto de análise de dados de doenças familiares (Bonney, 1986). No entanto, eles podem ser aplicados em várias áreas. O modelo logístico regressivo é utilizado principalmente por epidemiologistas como uma ferramenta estatística para estudar doenças familiares e outros traços binários. Nesse caso, os epidemiologistas desejam saber a significância da descendência como um fator de risco para a doença (Bonney, 1986). Por outro lado, os geneticistas preferem utilizar modelos de segregação com o objetivo de delinear a biologia do problema, ou seja, eles desejam clarear o modo de transmissão do gene através da segregação [1] das famílias em estudo e da sua associação e ligação com marcadores genéticos (Ott, 1999).

O intuito de Bonney (1986) foi misturar os objetivos dos epidemiologistas e dos geneticistas para estudar as doenças familiares e outros tratamentos binários.

O desenvolvimento de modelos para analisar respostas binárias dependentes com e sem utilização de variáveis explicativas foi apresentado em Bonney (1987). A parametrização consiste em decompor as probabilidades conjuntas em um produto de probabilidades condicionais.

[1] Transmissão de caracteres paternos para seus descendentes.

Neste capítulo, apresentaremos uma revisão de regressão logística e depois o modelo logístico regressivo.

2.2 Modelo de Regressão Logística

Suponha que desejamos estudar a relação entre uma determinada doença e seus fatores de risco. Nesse caso, a resposta estudada é a presença ou não da doença no indivíduo, e os fatores de risco podem ser, por exemplo, sexo, raça, idade etc. Para analisar esse tipo de situação, em geral, utiliza-se o modelo de regressão logística. Este tipo de modelo também é muito utilizado, por exemplo, em ensaios clínicos para comparação de tratamentos e na área financeira para classificação de clientes.

O modelo de regressão logística **dicotômico** é aquele que tem a variável resposta binária e que tem uma ou mais variáveis explicativas que podem ser tanto discretas como contínuas.

Seja Y uma variável resposta binária, ou seja,

$$P(Y = 1) = \pi \quad e \quad P(Y = 0) = 1 - \pi.$$

Se Y é uma variável aleatória com distribuição de Bernoulli, a média de Y é dada por,

$$E(Y) = 1 \times P(Y = 1) + 0 \times P(Y = 0) = P(Y = 1) = \pi. \tag{2.1}$$

Denote a probabilidade em (2.1) por $\pi(x)$, refletindo a dependência em valores das variáveis explicativas $\mathbf{X} = (X_1, ..., X_p)$. Então, a variância de Y é dada por,

$$V(Y) = E(Y^2) - [E(Y)]^2 = \pi(x)\,[1 - \pi(x)].$$

Se quisermos modelar Y como função de $\mathbf{X}$, teremos E(Y|x) = $\pi(x)$ e a função

de regressão logística é dada por,

$$\pi(x) \;=\; E(Y|x) \;=\; \frac{e^{(\beta_0 \,+\, \beta_1 \, x)}}{1 + e^{(\beta_0 \,+\, \beta_1 \, x)}}.$$

A função de ligação ou logito é dada por,

$$g(x) = log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x.$$

Considerando um modelo com p variáveis explicativas, a função de ligação tem a forma,

$$g(x) = log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 \, x_1 + \beta_2 \, x_2 + ... + \beta_p \, x_p \; .$$

2.2.1 Ajuste do Modelo Logístico por Máxima Verossimilhança

Seja $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ um vetor com n variáveis aleatórias independentes e identicamente distribuídas com distribuição de Bernoulli, ou seja, $Y_i \sim \text{Ber}(\pi)$, $i = 1, ..., n$. Seja x_i=(x_{i0}, x_{i1}, ..., x_{ip}) os i-ésimos valores de p variáveis explicativas, em que $x_{i0} = 1$. Pode-se escrever o modelo de regressão logística como

$$\pi(x_i) \;=\; \frac{exp\left(\sum_{j=0}^{p} \beta_j \, x_{ij}\right)}{\left[1 + exp\left(\sum_{j=0}^{p} \beta_j \, x_{ij}\right)\right]}. \tag{2.2}$$

Seja $V = \sum_{i=1}^{n} Y_i$ o número de "sucessos"(Y = 1) em n respostas binárias individuais, ou seja, V $\sim$ Bin(n, π). A função de probabilidade conjunta de (Y_1, ...,

Y_n) é igual ao produto de n funções de bernoulli

$$\begin{aligned} P(\mathbf{Y}) &= \prod_{i=1}^{n} P(Y_i = y_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \\ &= \left(\prod_{i=1}^{n}[1-\pi(x_i)]\right)\left(\prod_{i=1}^{n} exp\left[log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)^{y_i}\right]\right) \\ &= \left(\prod_{i=1}^{n}[1-\pi(x_i)]\right) exp\left[\sum_{i=1}^{n} y_i\ log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)\right]. \end{aligned} \tag{2.3}$$

Considerando que o i-ésimo logito é $\sum_{j=0}^{p} \beta_j\ x_{ij}$, então o termo exponencial na última expressão de (2.3) resulta em,

$$exp\left[\sum_{i=1}^{n} y_i \left(\sum_{j=0}^{p} \beta_j\ x_{ij}\right)\right] = exp\left[\sum_{j=0}^{p}\left(\sum_{i=1}^{n} y_i\ x_{ij}\right)\beta_j\right],$$

e, considerando que [1 - $\pi(x_i)$] = $\left[1 + exp\left(\sum_{j=0}^{p} \beta_j\ x_{ij}\right)\right]^{-1}$, a equação de log verossimilhança é dada por,

$$\log L(\boldsymbol{\beta}) = \sum_{j=0}^{p}\left(\sum_{i=1}^{n} y_i\ x_{ij}\right)\beta_j - \sum_{i=1}^{n}\ log\left[1 + exp\left(\sum_{j=0}^{p}\beta_j\ x_{ij}\right)\right].$$

Pode-se derivar as equações de verossimilhança diferenciando log L com respeito aos elementos de $\boldsymbol{\beta}$ e igualando-as à zero.

$$\frac{\partial logL}{\partial \beta_a} = \sum_{i=1}^{n} y_i\ x_{ia} - \sum_{i=1}^{n}\ x_{ia}\left[\frac{exp\left(\sum_{j=0}^{p}\beta_j\ x_{ij}\right)}{1 + exp\left(\sum_{j=0}^{p}\beta_j\ x_{ij}\right)}\right].$$

Então, as equações de verossimilhança são,

$$\sum_{i=1}^{n} y_i\ x_{ia} - \sum_{i=1}^{n}\ \hat{\pi}(x_i)\ x_{ia} = 0,\ \ a = 0, ..., p.$$

em que $\hat{\pi}(x_i)$ denota o estimador de máxima verossimilhança de $\pi(x_i)$, que é dado

por,

$$\hat{\pi}(x_i) = \frac{exp\left(\sum_{j=0}^{p} \hat{\beta}_j \; x_{ij}\right)}{1 + exp\left(\sum_{j=0}^{p} \hat{\beta}_j \; x_{ij}\right)}.$$

A matriz de informação de Fisher é o negativo do valor esperado da matriz de segundas derivadas parciais da log verossimilhança, ou seja,

$$I_F = -E\left[\frac{\partial^2 logL}{\partial \boldsymbol{\beta} \; \partial \boldsymbol{\beta^T}}\right].$$

Sob certas condições de regularidade (pg 126, Bickel), os estimadores de máxima verossimilhança dos parâmetros tem distribuição assintoticamente normal com matriz de covariância igual ao inverso da matriz de informação, ou seja, $\hat{\beta} \sim N(\beta, I_F^{-1})$. Para o modelo de regressão logística têm-se,

$$\begin{aligned} \frac{\partial^2 logL}{\partial \beta_a \partial \beta_b} &= -\sum_{i=1}^{n} \frac{x_{ia} \; x_{ib} \;\; exp\left(\sum_{j=0}^{p} \beta_j \; x_{ij}\right)}{\left[1 + exp\left(\sum_{j=0}^{p} \beta_j \; x_{ij}\right)\right]^2} \\ &= -\sum_{i=1}^{n} x_{ia} \; x_{ib} \;\; \pi_i \; (1 - \pi(x_i)). \end{aligned} \tag{2.4}$$

Como (2.4) não é uma função de $\{y_i\}$, as matrizes de segunda derivadas observadas e esperadas são idênticas.

A função de log verossimilhança para os modelos de regressão logística é estritamente côncava e as estimativas de máxima verossimilhança existem e são únicas exceto em alguns casos de fronteira. No entanto, as equações de verossimilhança são funções não-lineares dos estimadores de máxima verossimilhança de β, então para a obtenção dos estimadores é necessária uma solução iterativa. De acordo com Hosmer e Lemeshow (1989), o método iterativo mais utilizado é o de **Newton Raphson**.

O modelo de regressão logística com mais de duas categorias de resposta é chamado de modelo **politômico**. O modelo logístico politômico pode ser nominal ou ordinal, ou seja, se a variável resposta tem mais de duas categorias e essas

possuem uma ordem específica (por exemplo, ótimo, bom, regular e ruim), então o modelo é chamado de **politômico ordinal**, já se a variável resposta tem mais de duas categorias, e estas são nominais (não se pode assumir uma ordem para estas categorias), o modelo é **politômico nominal**. A modelagem é a mesma, independente da resposta ser nominal ou ordinal. Para maiores detalhes sobre modelos logísticos politômicos, veja Agresti (1990).

Quando estamos trabalhando com variável resposta categorizada, sabemos que um modelo adequado para esse tipo de variável resposta é o modelo de regressão logística, porém, assumimos independência entre as observações (indivíduos). No caso de dados de família essa suposição não é satisfeita, pois existe uma correlação entre os indivíduos da mesma família. Para solucionar esse tipo de problema, serão apresentados modelos alternativos que consideram alguma estrutura de dependência, como por exemplo, os modelos logísticos regressivos.

2.3 Modelos Logísticos Regressivos

Em estudos epidemiológicos, a influência de fatores genéticos em uma determinada doença é de grande interesse. Quando temos dados de família, em geral, os dados consistem de um número considerável de famílias nucleares [2] ou heredogramas [3]. Uma das dificuldades de analisar dados de família é o fato de que não podemos assumir independência dos indivíduos dentro de uma mesma família. Quando queremos estudar uma determinada doença e fatores (genéticos ou não) relacionados à essa doença utilizando as famílias como fonte de informações, devemos utilizar modelos que não suponham independência entre os membros da família. Assim, surge a necessidade de desenvolver novas metodologias estatísticas para análise desse tipo de dados. O desenvolvimento desses métodos estatísticos têm progredido muito ultimamente e muitos métodos têm sido apresentados, como por exemplo, os Modelos

[2] Familias Nucleares são famílias formadas por pai, mãe e filhos.

[3] Heredograma é uma representação gráfica de uma família, podendo incluir características como fenótipos e genótipos.

Logísticos Regressivos e muitas variações desses modelos utilizados em análise de segregação (Elston, 1980).

A Análise de Segregação é uma ferramenta utilizada para estudar dados familiares com a finalidade de estabelecer o modo de herança de uma determinada característica, quando o efeito de um gene não pode ser medido diretamente. Além disso, o objetivo da análise de segregação é detectar e estimar os efeitos de genes individuais no traço dentro de uma amostra de famílias. Especificamente, deseja-se estimar o número de alelos do gene que afeta o valor fenotípico, a freqüência dos alelos e a relação entre genótipo (conjunto dos genes de um indivíduo) e fenótipo. Um fenótipo é qualquer característica detectável de um organismo determinada pela interação entre o seu genótipo e o meio, como por exemplo, a cor dos cabelos ou dos olhos de um indivíduo.

Na análise de segregação, a covariância entre genótipos para um dado par é estimada usando não apenas o grau de relação, mas também os valores fenotípicos dos outros parentes.

Quando estamos analisando dados de família, a análise de segregação é o primeiro passo para determinar como um dado fenótipo foi herdado. A análise de segregação deseja discriminar entre os fatores que podem causar dependência entre os membros de uma família, ou seja, fatores genéticos, convívio compartilhado e hábitos culturais, para primeiramente testar a existência de um único gene, chamado de gene principal. O gene principal não é o único gene envolvido na expressão do fenótipo, embora, de todos os genes envolvidos, ele seja o que tem um efeito suficientemente importante para distingui-lo dos outros. Para entender exatamente quem são os genes principais, considere um exemplo simples: um locus autossômico[4] com dois alelos A e a. Então, os genes principais são: AA, Aa e aa.

Para um fenótipo clínico binário (afetado/não-afetado pela doença) esse efeito pode ser expresso em termos do risco relativo, por exemplo, a razão entre a probabilidade de ser afetado dado um genótipo AA e a probabilidade de ser afetado dado

[4]locus pertencente a um cromossomo não-sexual.

um genótipo *aa*. Para um fenótipo quantitativo, esse efeito é medido pela proporção da variância fenotípica explicada pelo gene principal (herdabilidade devido ao gene).

Em análise de segregação, os modelos regressivos e os modelos mistos fornecem resultados similares quando as correlações fenotípicas são devido à genes principais, e à fatores poligênicos aditivos (Bonney, 1984). No entanto, Bonney (1984, 1986) introduziu os modelos regressivos como uma alternativa aos modelos mistos utilizados em análise de segregação, já que os cálculos sob o modelo regressivo são mais simples do que sob os modelos mistos.

Bonney (1984) sugeriu alguns modelos regressivos para traços contínuos em famílias humanas. A idéia desses modelos é que a variável resposta para cada indivíduo é condicionada ao fenótipo desse indivíduo e no de seus ancestrais. Estes modelos podem avaliar, através de testes estatísticos, se o fenômeno (doença) tem efeito familiar ou não.

Bonney (1984) descreve classes naturais de modelos regressivos (Classe A, Classe B e Classe C) que utilizam estruturas Markovianas de dependência para analisar traços contínuos. Os modelos propostos utilizam estrutura de dependência de Markov entre o gene principal de um indivíduo e os genes principais de seus ancestrais.

Dados os genes principais dos pais, os genes da prole são independentes. Essa é a base dos modelos regressivos para a distribuição de genótipos em heredogramas. É também a base dos modelos Classe A para a distribuição dos resíduos sobre os heredogramas. Bonney (1984) propõe e discute modelos regressivos que podem ser utilizados para descrever covariação dentre e entre irmandades.

Bonney et al. (1988) utilizam modelos regressivos em análise de segregação e estendem esses modelos para incluir dados multivariados e loci de marcadores ligados.

Os modelos regressivos descrevem a dependência familiar especificando: (1) uma relação de regressão entre o fenótipo da pessoa e seu genótipo, (2) os fenótipos dos

pais e irmãos mais velhos e (3) outras variáveis explicativas (Bonney et al. 1988).

Utilizando os modelos regressivos, pode-se controlar variáveis epidemiologicamente importantes incorporando variáveis regressoras diretemente na análise de segregação. Os modelos regressivos propostos por Bonney (1984) utilizam estruturas Markovianas de dependência ao longo de linhas verticais de ancestrais, ou seja, a correlação entre avós e netos é o quadrado da correlação entre pais e filhos.

2.3.1 Modelos para Observações Binárias Dependentes

Como a probabilidade de uma observação é condicional em todas as observações precedentes, o modelo é chamado de modelo logístico regressivo (Bonney, 1987). Mesmo quando o modelo tem resposta binária dependente, o modelo logístico regressivo pode ser utilizado para descrever a dependência em termos das probabilidades condicionais.

A regressão logística é muito utilizada para estudar os efeitos de variáveis explicativas em respostas binárias. Para observações independentes, apenas um modelo logístico tem sido considerado, no entanto, vários modelos diferentes tem sido considerados para respostas dependentes (Cox, 1972).

A abordagem mais utilizada usa o modelo log-linear, que é dado por:

$$logP(\mathbf{Y}|\mathbf{X}) \;=\; \mu + \alpha_1 Z_1 + ... + \alpha_n Z_n + \alpha_{12} Z_1 Z_2 + ... + \alpha_{12...n} Z_1 Z_2 ... Z_n + \beta' X$$

em que os α's e β's são parâmetros e μ é a constante normalizadora escolhida para fazer as probabilidades $P(\mathbf{Y}|\mathbf{X})$ somarem um. Os Z's serão definidos em (2.8).

Muitos outros modelos para analisar respostas dicotômicas dependentes foram propostos. Stiratelli, et al. (1984) propuseram generalizações do modelo log-linear no caso de α e β serem conjuntamente normais bivariados e utilizando variáveis explicativas. Rosner (1984) converte o problema para uma regressão logística 2^n politômica assumindo uma distribuição beta-binomial para o número de sucessos quando não existem variáveis explicativas. No entanto, esses modelos têm alcance

limitado, pois todos descrevem apenas um padrão de dependência, chamado de correlação igual de Y's dado X's. Esses modelos também são limitados computacionalmente.

Bahadur (1961) sugeriu uma representação que utiliza 2^n - n - 1 parâmetros de correlação. Ou seja, considere $\boldsymbol{y} = (y_1 \ ... \ y_n)$ um vetor $n \times 1$ representando n respostas binárias. Cada y_i assume valor 1 ou 0, ou seja, a variável aleatória Y_i tem distribuição Bernoulli com parâmetro $\xi_i = \mathrm{P}\{Y_i = 1\}$. Então, $\mathrm{E}[Y_i] = \xi_i$ e $\mathrm{Var}[Y_i] = \xi_i(1 - \xi_i)$. Considere

$$Z_i = \frac{Y_i - \xi_i}{\sqrt{\xi_i(1-\xi_i)}}. \tag{2.5}$$

As correlações são dadas por,

$$\begin{aligned} r_{ij} &= E[Z_i Z_j], \\ r_{i\,j\,l} &= E[Z_i Z_j Z_l], \ ..., \ r_{12...n} = E[Z_1 Z_2 ... Z_n]. \end{aligned}$$

Assim, r_{ij} são as correlações de segunda ordem, $r_{i\,j\,l}$ são as correlações de terceira ordem e assim por diante. A desvantagem desse modelo é que ele não permite a inclusão de covariáveis.

Vamos agora apresentar os modelos logísticos regressivos no contexto de dados de família, descritos por Bonney (1986).

Considere um conjunto de n parentes com resposta binária para uma determinada doença, $\mathbf{Y} = (Y_1, ..., Y_n)$ em que $\mathbf{Y}$ é um vetor ordenado em ordem cronológica e é codificado da seguinte maneira:

$$Y_i = \begin{cases} 1 & \text{se o indivíduo } i \text{ tem a doença} \\ 0 & \text{caso contrário.} \end{cases} \tag{2.6}$$

Seja $\mathbf{X} = (X_1, ..., X_n)$ um vetor representando a variável explicativa, e assuma que o valor de X_i é associado com a resposta Y_i. O problema básico é expressar em termos significativos a probabilidade de $\mathbf{Y}$, ou seja, a probabilidade do indivíduo ter

a doença, como uma função da variável explicativa, $\mathbf{X}$, sem assumir independência entre os Y_i's. A abordagem regressiva para especificar um modelo logístico segue a seguinte decomposição da probabilidade de $\mathbf{Y}$ dado $\mathbf{X}$ em um produto de n probabilidades:

$$\begin{aligned} P(\mathbf{Y}|\mathbf{X}) &= P(Y_1,\ ...\ ,\ Y_n|\mathbf{X}) \\ &= P(Y_1|\mathbf{X})\ P(Y_2|Y_1,\mathbf{X})\ ...\ P(Y_n|Y_1,Y_2,\ ...\ ,\ Y_{n-1},\mathbf{X}). \end{aligned} \tag{2.7}$$

Então, o modelo logístico regressivo define uma função logística para cada fator em (2.7). Assuma também que

$$P(Y_i|\ Y_1,Y_2,\ ...\ ,\ Y_{i-1},\mathbf{X}) \ = \ P(Y_i|\ Y_1,Y_2,\ ...\ ,\ Y_{i-1},X_i).$$

Pode-se definir o i-ésimo logito da seguinte maneira:

$$\theta_i \ = \ log\left[\frac{P(Y_i=1|Y_1,...,Y_{i-1},X_i)}{P(Y_i=0|Y_1,...,Y_{i-1},X_i)}\right].$$

e assuma que θ_i é uma função linear de Y_1, ..., Y_{i-1}, X_i. Então, temos um problema de regressão no qual a resposta Y_i é binária, mas o conjunto dos valores da variável explicativa muda de acordo com i.

Será necessário criar variáveis auxiliares para introduzir dependência no modelo, ou seja, considere $Z_i = Z_i(Y_i)$. Assim, as variáveis Z_i's são funções lineares dos Y_i's dadas por:

$$Z_i = \begin{cases} 2Y_i - 1 & \text{se } Y_i = 0,\ 1, \\ 0 & \text{se } Y_i \text{ é } \textit{missing}. \end{cases} \tag{2.8}$$

Então, Z_i assume valores -1, 0, 1. Agora, definamos n logitos da seguinte forma:

$$\begin{aligned} \theta_1 &= \alpha + \beta X_1 \\ \theta_i &= \alpha + \sum_{j=1}^{i-1} \gamma_j\ Z_j + \beta\ X_i, \quad i\ =\ 2,...,n. \end{aligned} \tag{2.9}$$

em que α, β e γ's são parâmetros que variam no intervalo (-∞,+∞). A dependência foi introduzida no modelo através das variáveis Z_i's presentes nos logitos. No entanto, como θ_i é linear em Y_1, ..., Y_{i-1}, X_i, podemos considerar ao invés, a regressão da mesma resposta binária, Y_i, em Z_{i1}, ..., $Z_{i,i-1}$, X_i, em que $Z_{ij} = Z_{ij}(Y_j)$ são funções lineares conhecidas dos Y's. Para qualquer escolha dos Z's, os logitos definidos em (2.9) podem ser reescritos como,

$$\begin{aligned} \theta_i &= \alpha + \sum_{j=1}^{i-1} \gamma_j \, Z_{ij} + \beta X_i \\ &= \alpha + \gamma_1 \, Z_{i1} + ... + \gamma_{n-1} \, Z_{i,n-1} + \beta \, X_i. \end{aligned} \tag{2.10}$$

em que $Z_{ij} = 0$ para $j \geq i$. A equação (2.8) pode então ser reescrita como,

$$Z_{ij} = \begin{cases} 2Y_j - 1 & \text{se } j < i, \\ 0 & \text{se } j \geq i. \end{cases} \tag{2.11}$$

Então, (2.7) se torna,

$$P(\mathbf{Y}|\mathbf{X}) = \prod_{i=1}^{n} \frac{e^{\theta_i Y_i}}{(1 + e^{\theta_i})}, \tag{2.12}$$

em que o produtório é sobre os Y's observados. Sabe-se que se $Y_j = 1$ ($j < i$) a chance do indivíduo i ter a doença ($Y_i = 1$) aumenta em e^{γ_j}; Y_j desconhecido ou *missing* não muda a chance; $Y_j = 0$ diminui a chance em e^{γ_j} e um aumento de uma unidade em X_i aumenta a chance do indivíduo i ter a doença em e^{β}.

Colocando o modelo de forma matricial, tem-se:

$$\boldsymbol{\theta} = [\theta_1 \;\; \theta_2 \;\; ... \;\; \theta_n]',$$

$$\mathbf{Z} = [Z_1 \;\; Z_2 \;\; ... \;\; Z_n]',$$

$$\boldsymbol{\lambda} = [\alpha \;\; \gamma_1 \;\; \gamma_2 \;\; ... \;\; \gamma_{n-1} \;\; \beta]',$$

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & X_1 \\ 1 & Z_1 & 0 & \dots & 0 & X_2 \\ 1 & Z_1 & Z_2 & \dots & 0 & X_3 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & Z_1 & Z_2 & \dots & Z_{n-1} & X_n \end{pmatrix}.$$

Então, o modelo em (2.9) torna-se

$$\boldsymbol{\theta} \;=\; \mathbf{A}\,\boldsymbol{\lambda}.$$

Agora, pode-se olhar as colunas da matriz $\mathbf{A}$ como sendo variáveis explicativas para as respostas binárias $\mathbf{Y}$. Então, o modelo logístico regressivo (2.12) é formalmente equivalente a um modelo de regressão logística para n observações independentes.

A função de verossimilhança para n observações dependentes pode agora ser baseada em (2.12). Utilizando a notação matricial, o modelo para todas as n observações dependentes pode ser reescrito utilizando a notação descrita em (2.11) e (2.10), tendo a matriz $\mathbf{A}$ como:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & X_1 \\ 1 & Z_{21} & 0 & \dots & 0 & X_2 \\ 1 & Z_{31} & Z_{32} & \dots & 0 & X_3 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & Z_{n1} & Z_{n2} & \dots & Z_{n,n-1} & X_n \end{pmatrix} = \begin{pmatrix} 1 & Z_{11} & Z_{12} & \dots & Z_{1,n-1} & X_1 \\ 1 & Z_{21} & Z_{22} & \dots & Z_{2,n-1} & X_2 \\ 1 & Z_{31} & Z_{32} & \dots & Z_{3,n-1} & X_3 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & Z_{n1} & Z_{n2} & \dots & Z_{n,n-1} & X_n \end{pmatrix}.$$

Fica claro que o modelo logístico regressivo para n observações dependentes pode ser ajustado de maneira análoga ao modelo de regressão logística para n observações independentes.

A seguir, serão apresentados alguns modelos regressivos que não dependem de relações biológicas.

- Modelos Igualmente Preditivos

Considerando que $\gamma_1 = \gamma_2 = ... \ \gamma_{n-1} = \gamma$ nos logitos definidos em (2.9), tem-se que Y_1, Y_2, ..., Y_{i-1} tem valor preditivo aditivo e igual para Y_i. Denotemos a k-ésima soma parcial dos Z's por S_k:

$$S_k = \sum_{j=1}^{k} Z_j \ .$$

então, o logito dado em (2.9) torna-se:

$$\theta_i \ = \ \alpha + \gamma \ S_{i-1} + \beta X_i, \quad i \ = \ 2, ..., n. \tag{2.13}$$

A matriz **A** deve ser modificada de acordo com o novo modelo.

Para respostas igualmente preditivas, sabe-se que $Y_j = 1$ $(j < i)$ aumenta o logito da probabilidade de que $Y_i = 1$ em γ; $Y_j = 0$ reduz o logito em γ.

Seja S_k^+ o número de 1's ao longo das primeiras k respostas e S_k^- o número de zeros tal que $S_k = S_k^+$ - S_k^-. Outro modelo alternativo é, em (2.9) definir

$$\gamma_j = \begin{cases} \gamma^+ & \text{se } Z_j = 1, \\ \gamma^- & \text{se } Z_j = \text{-1}. \end{cases}$$

Então, uma generalização de (2.13) é dada por,

$$\theta_i \ = \ \alpha \ + \ \gamma^+ \ S_{i-1}^+ \ - \ \gamma^- \ S_{i-1}^- \ + \ \beta \ X_i. \tag{2.14}$$

Note que γ^- pode ser negativo para permitir que uma resposta negativa precedente tenha um efeito positivo na resposta corrente. O modelo (2.13) é um caso particular de (2.14), em que $\gamma^+ = \gamma^- = \gamma$.

- Estruturas Markovianas de Dependência

Quando seqüências de observações estão sendo estudadas, uma estrutura Markoviana simples de dependência geralmente descreve adequadamente a dependência

estocástica dos dados. Um modelo simples e bem conhecido para dependência serial tem uma estrutura de Markov de primeira ordem. Assim, (2.7) torna-se:

$$P(\mathbf{Y}|\mathbf{X}) \;=\; P(Y_1|\mathbf{X}) \prod_{i=2}^{n} P(Y_i|\, Y_{i-1},\, \mathbf{X}).$$

Assim, a probabilidade do indivíduo i ter a doença só depende da resposta do indivíduo imediatamente anterior a ele. Nesse caso, os logitos em (2.9) podem ser escritos como:

$$\theta_i \;=\; \alpha + \gamma Z_{i-1} + \beta X_i. \tag{2.15}$$

A modificação da matriz $\mathbf{A}$ e a extensão de (2.15) para estruturas Markovianas de ordem superior é imediata.

- Modelos Dependentes de Relações Biológicas

Serão definidos agora quatro tipos de modelos que utilizam as relações biológicas entre os membros da família. Os três primeiros falam sobre traços qualitativos analogamente aos modelos Classe A, Classe B e Classe C de Bonney (1984) para traços contínuos. É suposto que as observações são ordenadas tal que os pais precedam a prole e que os irmãos são ordenados de acordo com o nascimento. As letras C, F e M serão utilizadas para denotar cônjuge, pai e mãe do indivíduo i, respectivamente. B(1) e B(-1) denotam o irmão mais velho e o irmão imediatamente anterior, respectivamente, do indivíduo i.

Classe A

Essa é a classe mais simples de modelos para dependência dentro de irmandade. O modelo Classe A é aplicado quando as fontes de dependência entre os irmãos são: trasmissão biológica, transmissão cultural e convívio compartilhado com os pais (Bonney, 1984).

Esse modelo assume que, dado as respostas dos pais do indivíduo i, as respostas dos irmãos do indivíduo i não contribuem com mais informações em Y_i, ou seja,

$$P(Y_i|Y_1, Y_2, ..., Y_{i-1}, X_i) = P(Y_i|Y_F, Y_M, X_i). \quad (2.16)$$

No entanto, se o indivíduo i tem um cônjuge precedendo-o nos dados, e se as respostas do cônjuge são assumidas como correlacionadas, então Y_C, que é a resposta do cônjuge de i, também tem informação em Y_i. Então, um modelo mais geral seria,

$$P(Y_i|Y_1, Y_2, ..., Y_{i-1}, X_i) = P(Y_i|Y_C, Y_F, Y_M, X_i). \quad (2.17)$$

Então, os logitos em (2.9) seguem a seguinte forma,

$$\theta_i = \alpha + \gamma_C\, Z_C + \gamma_F\, Z_F + \gamma_M\, Z_M + \beta\, X_i. \quad (2.18)$$

Note que, por definição, $Z_C = 0$ se o indivíduo i não tem cônjuge, ou se o cônjuge não o precede nos dados. Analogamente, se um dos pais não consta nos dados, o Z correspondente a esse pai é zero.

Classe B

Esse modelo generaliza os modelos Classe A, pois este assume que, dado as respostas dos pais e dos r primeiros filhos, as respostas dos filhos restantes são independentes. Para r = 1, têm-se,

$$P(Y_i|Y_1, Y_2, ..., Y_{i-1}, X) = P(Y_i|Y_C, Y_F, Y_M, Y_{B(1)}, X_i). \quad (2.19)$$

$$\theta_i = \alpha + \gamma_C\, Z_C + \gamma_F\, Z_F + \gamma_M\, Z_M + \gamma_{B(1)}\, Z_{B(1)} + \beta\, X_i.$$

Note que se o indivíduo i é o filho mais velho, então, por definição B(1) é *missing*, e $Z_{B(1)} = 0$. Para r = 2, têm-se:

$$P(Y_i|Y_1, Y_2, ..., Y_{i-1}, X) = P(Y_i|Y_C, Y_F, Y_M, Y_{B(1)}, Y_{B(2)}X_i). \quad (2.20)$$

em que B(2) representa o segundo filho do casal.

$$\theta_i \;=\; \alpha + \gamma_C\, Z_C + \gamma_F\, Z_F + \gamma_M\, Z_M + \gamma_{B(1)}\, Z_{B(1)} + \gamma_{B(2)}\, Z_{B(2)} + \beta\, X_i.$$

Note que se o indivíduo i é o segundo filho do casal, então, temos o caso em que r=1 e por definição B(2) é *missing*, e $Z_{B(2)} = 0$.

Classe C

Esse modelo segue a intuição natural de que, em uma grande irmandade, assume-se que irmãos que nascem próximos são mais correlacionados do que irmãos que nascem com uma distância maior um do outro. Assim, a resposta do indivíduo i depende da resposta dos pais, do cônjuge e dos r irmãos imediatamente anteriores a ele. Então, assumindo uma dependência serial entre os irmãos e r=1, têm-se:

$$P(Y_i|Y_1, Y_2, ..., Y_{i-1}, X) \;=\; P(Y_i|Y_C, Y_F, Y_M, Y_{B(-1)}, X_i), \qquad (2.21)$$

$$\theta_i \;=\; \alpha + \gamma_C\, Z_C + \gamma_F\, Z_F + \gamma_M\, Z_M + \gamma_{B(-1)}\, Z_{B(-1)} + \beta\, X_i.$$

Se o indivíduo i é o irmão mais velho, então, B(-1) é *missing* e $Z_{B(-1)} = 0$. Para r=2, têm-se:

$$P(Y_i|Y_1, Y_2, ..., Y_{i-1}, X) \;=\; P(Y_i|Y_C, Y_F, Y_M, Y_{B(-1)}, Y_{B(-2)}, X_i), \qquad (2.22)$$

$$\theta_i \;=\; \alpha + \gamma_C\, Z_C + \gamma_F\, Z_F + \gamma_M\, Z_M + \gamma_{B(-1)}\, Z_{B(-1)} + \gamma_{B(-2)}\, Z_{B(-2)} + \beta\, X_i,$$

em que B(-2) representa o segundo irmão imediatamente anterior ao indivíduo i. Se o indivíduo i é o segundo filho do casal, então, B(-2) é *missing* e $Z_{B(-2)} = 0$. Nesse caso, os modelos Classe B e Classe C são idênticos.

É importante notar que, se para um valor específico de r as principais suposições dos modelos Classe B ou Classe C não são consideradas realistas, essas suposições podem ser relaxadas igualando-se r ao tamanho da maior

irmandade (Bonney, 1984).

Seria interessante testar o modelo Classe A contra os modelos alternativos Classe B ou Classe C. Dentre os modelos Classe B ou Classe C, testes devem ser feitos para determinar a ordem mínima de dependência (r) dentro das irmandades (Bonney, 1984).

Classe D

Esse modelo assume que dado as respostas dos pais, as respostas da prole são igualmente preditivas. Seja S_{os} a soma sobre os irmãos mais velhos, então os logitos em (2.9) devem ser escritos como,

$$\theta_i \;=\; \alpha + \gamma_C\, Z_C + \gamma_F\, Z_F + \gamma_M\, Z_M + \gamma\, Z_{S_{os}} + \beta\, X_i.$$

Uma versão composta dos modelos regressivos Classe D para p respostas fenotípicas pode ser vista em Bagchi et al. (1993).

Bonney (1992) apresenta uma extensão dos modelos regressivos para análise de dados de família, para incluir casos em que a covariação dentro da irmandade deve exceder aquela suposta pelos modelos regressivos Classe A, mas a ordem dos nascimentos não é necessária. Então, esse modelo regressivo composto é uma versão dos modelos regressivos Classe D com a propriedade de mudança dentro da irmandade.

2.3.2 Cálculo da Função de Verossimilhança

Para o modelo logístico dado em (2.12), a distribuição conjunta é um produto de funções logísticas univariadas:

$$P(\mathbf{Y}|\mathbf{X}) \;=\; \prod_{i=1}^{n} \frac{e^{\theta_i(\boldsymbol{\lambda})\, Y_i}}{(1 + e^{\theta_i(\boldsymbol{\lambda})})},$$

em que $\boldsymbol{\lambda}$=[α, ..., β]'.

A função de log-verossimilhança é dada por,

$$l = log\ L(\boldsymbol{\lambda}) = \sum_{i=1}^{n} log\ \left[\frac{e^{\theta_i(\boldsymbol{\lambda})\ Y_i}}{(1+e^{\theta_i(\boldsymbol{\lambda})})}\right] = \sum_{i=1}^{n} [Y_i\ \theta_i(\boldsymbol{\lambda}) - log(1+e^{\theta_i(\boldsymbol{\lambda})})].$$

Os estimadores de máxima verossimilhança de $\boldsymbol{\lambda}$ podem ser obtidos por meio das primeiras derivadas parciais da log verossimilhança, igualando-as a zero.

$$\frac{\partial\ logL}{\partial \boldsymbol{\lambda}} = \sum_{i=1}^{n} Y_i \times \left[\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\lambda}}\right] - \sum_{i=1}^{n} \frac{e^{\theta_i(\boldsymbol{\lambda})}}{log(1+e^{\theta_i(\boldsymbol{\lambda})})} \times \left[\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\lambda}}\right] \tag{2.23}$$

$$\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\lambda}} = \left[\frac{\partial \boldsymbol{\theta}}{\partial \alpha}\ \frac{\partial \boldsymbol{\theta}}{\partial \gamma_1}\ \ldots\ \frac{\partial \boldsymbol{\theta}}{\gamma_{n-1}}\ \frac{\partial \boldsymbol{\theta}}{\partial \beta}\right]'. \tag{2.24}$$

É necessário utilizar métodos iterativos para solucionar as equações, como por exemplo, o método de Newton Raphson.

Se denotarmos o estimador de máxima verossimilhança do k-ésimo parâmetro por $\hat{\lambda_k}$, a matriz de informação de Fisher, I_F, é usada para obter os erros padrão das estimativas dos parâmetros. A matriz de informação de Fisher observada é dada por:

$$I_F(\boldsymbol{\lambda}) = -\frac{\partial^2 logL(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}\ \partial \boldsymbol{\lambda}^T} = \sum_{i=1}^{n} \frac{e^{\theta_i(\boldsymbol{\lambda})}}{[1+e^{\theta_i(\boldsymbol{\lambda})}]^2}.\mathbf{V}_i\mathbf{V}_i^T, \tag{2.25}$$

em que $\mathbf{V}_i = \frac{\partial \theta_i(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}$.

Hipóteses podem ser testadas utilizando-se o teste da razão de verossimilhanças. Uma hipótese de interesse consiste em testar um modelo independente contra alguma suposta estrutura de dependência. Seria interessante comparar o ajuste de vários padrões de dependência que não precisam necessariamente estar aninhados como no teste da razão de verossimilhanças. Uma comparação rápida baseada na teoria de verossimilhança é obtida usando o critério de informação de Akaike (AIC).

O **AIC** é definido por,

$$AIC = -2 \times log(\text{verossimilhança}) + 2 \times (\text{número de parâmetros}).$$

Quanto mais parâmetros um modelo contém, menos acuradas são as estimativas desses parâmetros. Além disso, se forem ajustados mais parâmetros do que necessário, a qualidade das estimativas dos parâmetros cai. O AIC é um critério de seleção de modelos que leva em consideração o número de parâmetros estimados do modelo. A idéia básica é selecionar um modelo que seja parcimonioso, ou, em outras palavras, que esteja bem ajustado e tenha um número de parâmetros reduzido. Assim, o modelo com o menor AIC é o que ajusta melhor os dados. Esse critério tem sido considerado consistente para modelos não aninhados. Apesar do critério não mostrar que o melhor entre dois modelos é "significativamente melhor", ele é uma ferramenta útil de comparação rápida para modelos paramétricos.

3 Modelos para Analisar Seqüências de DNA

3.1 Introdução

Com o desenvolvimento rápido de técnicas de sequenciamento, foi necessário o desenvolvimento de programas de computador capazes de manipular, analisar e comparar longas seqüências de dados.

Muitos estudos foram feitos para analisar seqüências de DNA, no entanto, esses estudos consideravam que as bases na seqüência estavam uniformente dispostas. No entanto, os dados sugerem que dentre várias possíveis escolhas de códons para codificar um determinado aminoácido, existem preferências. Por exemplo, o aminoácido *glutamina* pode ser codificado pelos códons CAG ou CAA, mas o códon CAG ocorre com uma freqüência bem maior em alguns genes (Manistis et al., 1982).

Existe uma forte correlação entre as freqüências dos códons e a abundância relativa dos tRNA's correspondentes (Ikemura, 1981;1982). Os códons aparecem com uma freqüência desigual em seqüências codificadoras, assim, seria interessante estudar essas freqüências para analisar esse tipo de conjunto de dados.

A ordem das bases numa seqüência de DNA é influenciada por fatores aleatórios (mutações) e por pressões determinísticas (seleção). Seqüências de DNA de organismos procariontes e eucariontes apresentam padrões de vizinhança diferentes para as bases da seqüência.

A ordem linear das bases no DNA e a possível dependência estatística das bases de nucleotídeo em um códon específico faz com que o modelo logístico regressivo seja uma ferramenta apropriada para analisar as freqüências dos códons. Além disso, o mecanismo pelo qual as seqüências de DNA são produzidas, sequencialmente em longas cadeias, sugere que a análise dessas seqüências deve ser feita utilizando-se a metodologia das Cadeias de Markov.

A utilização de Cadeias de Markov na análise de seqüências de DNA tem sido uma ferramenta estatística muito utilizada (Bonney, 1987; Tavaré & Giddings, 1988). Esses modelos consistem em estudar a dependência entre os nucleotídeos usando modelos de Markov. Tavaré & Giddings (1988) apresentam a metodologia de Cadeias de Markov para analisar seqüências de DNA ou RNA. Além disso, eles tem por objetivo estimar a ordem da Cadeia de Markov, ou seja, eles determinam a distância na qual existe dependência de base modelando a seqüência como uma Cadeia de Markov.

Assim, seja X $= \{X_n, n = 1, 2, ...\}$ um processo estocástico com m estados. Se esses estados representam os nucleotídeos em uma dada seqüência de DNA, então têm-se $m = 4$ (A = 1, C = 2, G = 3 e T = 4). X é chamado de Cadeia de Markov de ordem k se

$$\begin{aligned} &P\{X_{n+1} = i_{n+1} | X_n = i_n, ..., X_{n-k+1} = i_{n-k+1}, ..., X_1 = i_i\} \\ &\quad = P\{X_{n+1} = i_{n+1} | X_n = i_n, ..., X_{n-k+1} = i_{n-k+1}\}, \end{aligned} \tag{3.1}$$

para todo $n > k-1$ e para qualquer escolha dos estados i_1, i_2,..., i_{n-1} de $\{1, 2, ..., m\}$. Em outras palavras, a distribuição da próxima base na seqüência é deternimada pelas k bases anteriores. Quando $k = 0$, as bases são distribuídas independentemente. A cadeia de Markov usual, chamada de Cadeia de Markov de primeira ordem, é obtida quando $k = 1$.

Uma análise típica para uma única seqüência seria testar a independência entre as bases, ou seja, testar se $k = 0$, testando-se assim a uniformidade da composição da base, testando o ajuste de uma matriz de transição hipotética e testando proba-

bilidades de transição particulares.

Alguns estudos de simulação limitados nas propriedades do modelo mostraram que ignorar a dependência estatística entre os nucleotídeos pode levar a estimativas viciadas dos efeitos das covariáveis (Bonney & Brooks, 1989). No entanto, a maioria desses estudos não consideram a possível influência de variáveis exógenas na freqüência de um códon. Um exemplo seria considerar uma variável que mede a propensão de ocorrer mutação entre um códon e um códon de parada. Outra variável correlacionada é o risco mutacional, que mede a possibilidade de um códon mudar de uma base única (códon com três bases iguais) para uma tripla (códon com as três bases diferentes) codificando um aminoácido muito diferente. Essas mutações mudariam o gene e poderiam causar prejuízos para o organismo.

3.2 Modelos Logísticos Regressivos para Dados Politômicos

Nosso principal interesse é a relação entre as freqüências dos códons em um dado gene e as possíveis variáveis explicativas. Como cada códon é uma tripla ordenada das bases A, C, G e T, considera-se uma resposta trivariada. Conceitualmente, os códons em um gene específico podem ser considerados como uma amostra selecionada, não necessariamente aleatória, de um conjunto infinito de triplas formadas pelas bases A, C, G e T. Seja B_1, B_2 e B_3 as três variáveis resposta que representam as bases que formam o códon, em que $B_i \in \{\text{A, C, G , T}\}$; i = 1, 2, 3.

Seja $\mathbf{X} = (X_1, X_2, ..., X_p)$ o vetor de variavéis explicativas. O problema estatístico é especificar a probabilidade de um códon ou, equivalentemente, a probabilidade conjunta de uma tripla ordenada de bases (B_1, B_2, B_3) em termos das covariáveis $\mathbf{X}$, ou seja, $P(B_1, B_2, B_3|\mathbf{X})$ considerando apenas uma seqüência de DNA.

Os códons de parada são códons que não codificam aminoácidos, eles simplesmente identificam o término da síntese protéica, então, seria interessante considerar

uma probabilidade normalizada com relação aos códons efetivos, ou seja, aqueles que realmente codificam aminoácidos. Assim, obtém-se,

$$W(B_1, B_2, B_3|\mathbf{X}) = \frac{P(B_1, B_2, B_3|\mathbf{X})}{\sum_\eta P(B_1, B_2, B_3|\mathbf{X})}, \tag{3.2}$$

em que η é a soma sobre os códons efetivos, isto é, excluindo os códons de parada. Como o códon é uma tripla ordenada das bases B_1, B_2, B_3, é apropriado usar modelos que levem em consideração a ordem.

Nesse caso, os modelos regressivos podem ser apropriados nesse contexto e serão dados pela decomposição das seguintes probabilidades:

$$P(B_1, B_2, B_3|\mathbf{X}) \quad em \quad P(B_1|\mathbf{X})P(B_2|B_1, \mathbf{X})P(B_3|B_1, B_2, \mathbf{X}), \tag{3.3}$$

especificando um modelo de regressão logística politômica para cada um dos três fatores. Como as bases do DNA (A, C, G e T) não possuem uma ordem específica, vamos considerar os modelos logísticos regressivos politômicos nominais.

As categorias de resposta são quatro (A, C, G e T), então têm-se três logitos para cada posição do códon(B_i). Assim, codificando as categorias de B_i como:

$$T = 0, \quad C = 1, \quad A = 2 \quad e \quad G = 3,$$

para o i-ésimo fator da decomposição, $i = 1, 2, 3$, pode-se escrever os logitos como:

$$\theta_{ij} = log\frac{P(B_i = j|B_1, ..., B_{i-1}, \mathbf{X})}{P(B_i = 0|B_1, ..., B_{i-1}, \mathbf{X})}, \quad i,\ j = 1, 2, 3.$$

Então, os fatores da decomposição são dados por:

$$P(B_i = 0|B_1, ..., B_{i-1}, \mathbf{X}) = \frac{1}{1 + \sum_{j=1}^{3} exp(\theta_{ij})} \quad e$$

$$P(B_i = j|B_1, ..., B_{i-1}, \mathbf{X}) = \frac{exp(\theta_{ij})}{1 + \sum_{j=1}^{3} exp(\theta_{ij})}.$$

Para a construção do modelo logístico regressivo, será necessário construir variáveis indicadoras para os B's, para introduzir a dependência no modelo, como

descrito na seção 2.31. Sejam:

$$Z_{i1} = \begin{cases} 1 & \text{se } B_i = \text{C (ou 1);} \\ 0 & \text{caso contrário;} \end{cases}$$

$$Z_{i2} = \begin{cases} 1 & \text{se } B_i = \text{A (ou 2);} \\ 0 & \text{caso contrário;} \end{cases}$$

$$Z_{i3} = \begin{cases} 1 & \text{se } B_i = \text{G (ou 3);} \\ 0 & \text{caso contrário;} \end{cases}$$

de tal forma que, se $Z_{i1} = Z_{i2} = Z_{i3} = 0$, então $B_i = \text{T}$ (casela de referência). A tabela abaixo mostra essa codificação.

Tabela 3.5: Codificação

Base B_i	Código	Variáveis Dummy		
		Z_{i1}	Z_{i2}	Z_{i3}
T	0	0	0	0
C	1	1	0	0
A	2	0	1	0
G	3	0	0	1

Serão estudados quatro modelos para analisar as freqüências dos códons na seqüência de DNA. No entanto, cada modelo será dividido em outros dois modelos, ou seja, modelo com covariáveis e modelo sem covariáveis, para estudarmos também a influência das covariáveis nas freqüências dos códons. Três desses modelos apresentam algum tipo de estrutura de dependência entre os nucleotídeos e o quarto modelo assume independência entre os nucleotídeos.

- Modelo Aditivo

  O modelo aditivo é definido da seguinte maneira:

$$\begin{aligned}
\theta_{1j} &= \alpha_{1j} + \sum_{k=1}^{p} \beta_k X_k; \\
\theta_{2j} &= \alpha_{2j} + \sum_{s=1}^{3} \gamma_{1s} Z_{1s} + \sum_{k=1}^{p} \beta_k X_k; \\
\theta_{3j} &= \alpha_{3j} + \sum_{s=1}^{3} [\gamma_{1s} Z_{1s} + \gamma_{2s} Z_{2s}] + \sum_{u=1}^{3} \sum_{v=1}^{3} \tau_{12(u,v)} Z_{1u} Z_{2v} \\
&+ \sum_{k=1}^{p} \beta_k X_k; \quad j = 1,\ 2\ e\ 3. \qquad (3.4)
\end{aligned}$$

  em que os $\alpha's$, $\gamma's$ e $\beta's$ são parâmetros desconhecidos. Os $\gamma's$ representam o efeito dos nucleotídeos precedentes na i-ésima resposta, os $\tau's$ representam as interações dos nucleotídeos com os parâmetros e os $\beta's$ representam os efeitos das covariavéis. Na realidade, esse modelo não é o mais geral, pois ele considera apenas interações de primeira ordem entre os nucleotídeos e não considera as interações entre as covariáveis, nem entre nucleotídeos e covariáveis. Note que esse modelo, mesmo sem todas as interações já apresenta um número muito grande de parâmetros.

- Modelos Igualmente Preditivos

  Considere a i-ésima posição em um códon e seja S_{i-1}^{C} o número de C's anteriores. De maneira análoga, defina $S_{i-1}^{T}, S_{i-1}^{A}, S_{i-1}^{G}$. Assim, pode-se definir um modelo igualmente preditivo no caso de $\gamma_{1s} = \gamma_{2s} = \gamma_s$ (Bonney, 1987), então:

$$\theta_{ij} = \alpha_{ij} + \gamma_1 S_{i-1}^{C} + \gamma_2 S_{i-1}^{A} + \gamma_3 S_{i-1}^{G} + \sum_{k=1}^{p} \beta_k X_k. \qquad (3.5)$$

  Considerando esse modelo, cada C anterior, por exemplo, aumenta o logito de A, C ou G pela mesma quantidade γ_1, cada A anterior aumenta o logito em γ_2 e cada G anterior aumenta o logito em γ_3. Então, para o terceiro logito o efeito preditivo de um G anterior não depende da posição em que G ocorreu,

na primeira ou segunda posição da tripla. No entanto, note que é possível que A, C e G tenham efeitos preditivos diferentes porque em geral γ_1, γ_2 e γ_3 podem ser diferentes. Indexando os $\gamma's$, obtém-se um modelo mais geral que (3.5):

$$\theta_{ij} \quad = \quad \alpha_{ij} + \gamma_{1j} S_{i-1}^{C} + \gamma_{2j} S_{i-1}^{A} + \gamma_{3j} S_{i-1}^{G} + \sum_{k=1}^{p} \beta_k X_k. \tag{3.6}$$

Para o modelo (3.6), cada C anterior, aumenta o logito de B_j em γ_{1j}, cada G anterior em γ_{3j} e assim por diante. Apenas nucleotídeos similares são igualmente preditivos nesse modelo. Pode-se ajustar um modelo com uma versão mais parcimoniosa dos modelos igualmente preditivos em que os $\gamma's$ não dependem de i e j, e nesse caso, os nucleotídeos anteriores, independente do tipo, são igualmente preditivos. Então, o modelo é dado por:

$$\theta_{ij} \quad = \quad \alpha_i + \gamma(S_{i-1}^{C} + S_{i-1}^{A} + S_{i-1}^{G}) + \sum_{k=1}^{p} \beta_k X_k. \tag{3.7}$$

Note que $S_{i-1}^{T} + S_{i-1}^{C} + S_{i-1}^{A} + S_{i-1}^{G} = i - 1$.

Note que esse modelo pode ser interpretado como uma regressão em série da ordem dos nucleotídeos anteriores que não são iguais a T. Note também que as variáveis explicativas linearmente relacionadas com a ordem em série i dos nucleotídeos não podem ser usadas como variáveis explicativas nos modelos igualmente preditivos se uma incorporar todos os quatro S's no modelo.

- Estruturas Markovianas de Dependência

  Uma outra simplificação da estrutura de dependência é dada pela decomposição:

  $$P(B_1, B_2, B_3|\mathbf{X}) = P(B_1|\mathbf{X})P(B_2|B_1, \mathbf{X})P(B_3|B_2, \mathbf{X}).$$

  Essa é a estrutura de Markov de ordem 1, onde os logitos dependem apenas do nucleotídeo imediatamente anterior. O modelo de Markov para os logitos j

= 1, 2, 3 é dado por:

$$\theta_{ij} = \alpha_{ij} + \sum_{s=1}^{3} \gamma_{is} Z_{i-1,s} + \sum_{k=1}^{p} \beta_k X_k.$$

Para cada i, o j-ésimo logito depende apenas o nucleotídeo anterior, mas o efeito preditivo de um C anterior não é o mesmo de um G anterior. Modelos mais simples, com um número menor de parâmetros são dados por,

$$\theta_{ij} = \alpha_i + \sum_{s=1}^{3} \gamma_{1s} Z_{i-1,s} + \sum_{k=1}^{p} \beta_k X_k;$$

$$\theta_{ij} = \alpha_i + \sum_{s=1}^{3} \gamma Z_{i-1,s} + \sum_{k=1}^{p} \beta_k X_k.$$

Esses modelos tem foco na regressão da resposta imediatamente anterior, no entanto, $Z_{i-1,s}$ deve ser trocada por $Z_{1,s}$ se existir razão para acreditar que é a primeira base, ao invés da base anterior, que influencia a i-ésima base.

- Modelo Independente

O modelo independente é o mais simples de todos os modelos que iremos estudar, pois ele não considera nenhuma estrutura de dependência entre os nucleotídeos. Assim, para o modelo com as covariáveis, têm-se:

$$\theta_{ij} \;=\; \alpha_{ij} + \sum_{k=1}^{p} \beta_k X_k; \qquad i,\ j \;=\; 1,\ 2\ e\ 3.$$

3.3 Verossimilhanças e Matriz de Informação

Denotemos a freqüência dos códons B_1, B_2, B_3 nos dados para uma medida $\mathbf{X}$ das covariáveis por $n(B_1, B_2, B_3|\mathbf{X})$, e assim, a verossimilhança dos dados para uma

seqüência de DNA pode ser baseada no modelo multinomial:

$$L \propto \prod_{B_1=0}^{3} \prod_{B_2=0}^{3} \prod_{B_3=0}^{3} W(B_1, B_2, B_3|\mathbf{X})^{n(B_1,B_2,B_3|\mathbf{x})}$$

em que $W(B_1, B_2, B_3|\mathbf{X})$ já foi definido em (3.2). O uso da verossimilhança multinomial na análise de seqüências de DNA não é um fato novo. A novidade aqui é o uso das covariáveis e do modelo logístico regressivo nesse contexto.

Considere o vetor de respostas $\mathbf{B} = (B_1,\ B_2,\ B_3)$, $B_i \in \{\text{T, C, A, G}\}$, em que i indica a posição no códon. Então, pode-se assumir que $B_i \sim \text{Mult}(M_i,\ p_{i1},\ p_{i2},\ p_{i3},\ p_{i4})$. Vamos apresentar os logitos para as três posições do códon para visualizar como a estrutura de dependência é implantada na verossimilhança.

Os logitos para a primeira posição do códon são dados por:

$$\begin{aligned}
\theta_{11} &= log\left[\frac{P(B_1 = 1|\mathbf{X})}{P(B_1 = 0|\mathbf{X})}\right] = log\left[\frac{P(B_1 = C|\mathbf{X})}{P(B_1 = T|\mathbf{X})}\right], \\
\theta_{12} &= log\left[\frac{P(B_1 = 2|\mathbf{X})}{P(B_1 = 0|\mathbf{X})}\right] = log\left[\frac{P(B_1 = A|\mathbf{X})}{P(B_1 = T|\mathbf{X})}\right] e \\
\theta_{13} &= log\left[\frac{P(B_1 = 3|\mathbf{X})}{P(B_1 = 0|\mathbf{X})}\right] = log\left[\frac{P(B_1 = G|\mathbf{X})}{P(B_1 = T|\mathbf{X})}\right].
\end{aligned}$$

Para a segunda posição do códon:

$$\begin{aligned}
\theta_{21} &= log\left[\frac{P(B_2 = 1|B_1, \mathbf{X})}{P(B_2 = 0|B_1, \mathbf{X})}\right] = log\left[\frac{P(B_2 = C|B_1, \mathbf{X})}{P(B_2 = T|B_1, \mathbf{X})}\right], \\
\theta_{22} &= log\left[\frac{P(B_2 = 2|B_1, \mathbf{X})}{P(B_2 = 0|B_1, \mathbf{X})}\right] = log\left[\frac{P(B_2 = A|B_1, \mathbf{X})}{P(B_2 = T|B_1, \mathbf{X})}\right] e \\
\theta_{23} &= log\left[\frac{P(B_2 = 3|B_1, \mathbf{X})}{P(B_2 = 0|B_1, \mathbf{X})}\right] = log\left[\frac{P(B_2 = G|B_1, \mathbf{X})}{P(B_2 = T|B_1, \mathbf{X})}\right].
\end{aligned}$$

Analisando os logitos para a segunda posição do códon é fácil ver que a base que aparece na segunda posição do códon depende da base que apareceu na primeira posição do códon, assim, estamos modelando a estrutura de dependência através

dos logitos. Para a terceira posição do códon, têm-se:

$$\theta_{31} = log\left[\frac{P(B_3 = 1|B_1, B_2, \mathbf{X})}{P(B_3 = 0|B_1, B_2, \mathbf{X})}\right] = log\left[\frac{P(B_3 = C|B_1, B_2, \mathbf{X})}{P(B_3 = T|B_1, B_2, \mathbf{X})}\right],$$
$$\theta_{32} = log\left[\frac{P(B_3 = 2|B_1, B_2, \mathbf{X})}{P(B_3 = 0|B_1, B_2, \ \mathbf{X})}\right] = log\left[\frac{P(B_3 = A|B_1, B_2, \mathbf{X})}{P(B_3 = T|B_1, B_2, \mathbf{X})}\right] e$$
$$\theta_{33} = log\left[\frac{P(B_3 = 3|B_1, B_2, \mathbf{X})}{P(B_3 = 0|B_1, B_2, \mathbf{X})}\right] = log\left[\frac{P(B_3 = G|B_1, B_2, \mathbf{X})}{P(B_3 = T|B_1, B_2, \mathbf{X})}\right].$$

De maneira análoga, os logitos da terceira posição também incluem estrutura de dependência entre as bases do códon.

As probabilidades da multinomial para a primeira posição do códon são dadas por:

$$p_{11} = P(B_1 = 1|\mathbf{X}) = P(B_1 = C|\mathbf{X}) = \frac{e^{\theta_{11}}}{1 + e^{\theta_{11}} + e^{\theta_{12}} + e^{\theta_{13}}}$$
$$p_{12} = P(B_1 = 2|\mathbf{X}) = P(B_1 = A|\mathbf{X}) = \frac{e^{\theta_{12}}}{1 + e^{\theta_{11}} + e^{\theta_{12}} + e^{\theta_{13}}}$$
$$p_{13} = P(B_1 = 3|\mathbf{X}) = P(B_1 = G|\mathbf{X}) = \frac{e^{\theta_{13}}}{1 + e^{\theta_{11}} + e^{\theta_{12}} + e^{\theta_{13}}}$$
$$p_{14} = P(B_1 = 0|\mathbf{X}) = P(B_1 = T|\mathbf{X}) = \frac{1}{1 + e^{\theta_{11}} + e^{\theta_{12}} + e^{\theta_{13}}},$$

pois $\sum_{j=1}^{4} p_{1j} = 1$.

As probabilidades da multinomial para a segunda posição do códon são dadas por:

$$p_{21} = P(B_2 = 1|\mathbf{Z},\ \mathbf{X}) = P(B_2 = C|\mathbf{Z},\ \mathbf{X}) = \frac{e^{\theta_{21}}}{1 + e^{\theta_{21}} + e^{\theta_{22}} + e^{\theta_{23}}}$$
$$p_{22} = P(B_2 = 2|\mathbf{Z},\ \mathbf{X}) = P(B_2 = A|\mathbf{Z},\ \mathbf{X}) = \frac{e^{\theta_{22}}}{1 + e^{\theta_{21}} + e^{\theta_{22}} + e^{\theta_{23}}}$$
$$p_{23} = P(B_2 = 3|\mathbf{Z},\ \mathbf{X}) = P(B_2 = G|\mathbf{Z},\ \mathbf{X}) = \frac{e^{\theta_{23}}}{1 + e^{\theta_{21}} + e^{\theta_{22}} + e^{\theta_{23}}}$$
$$p_{24} = P(B_2 = 0|\mathbf{Z},\ \mathbf{X}) = P(B_2 = T|\mathbf{Z},\ \mathbf{X}) = \frac{1}{1 + e^{\theta_{21}} + e^{\theta_{22}} + e^{\theta_{23}}},$$

pois $\sum_{j=1}^{4} p_{2j} = 1$.

As probabilidades para a terceira posição do códon são dadas por:

$$p_{31} = P(B_3 = 1|\mathbf{Z},\ \mathbf{X}) = P(B_3 = C|\mathbf{Z},\ \mathbf{X}) = \frac{e^{\theta_{31}}}{1 + e^{\theta_{31}} + e^{\theta_{32}} + e^{\theta_{33}}}$$
$$p_{32} = P(B_3 = 2|\mathbf{Z},\ \mathbf{X}) = P(B_3 = A|\mathbf{Z},\ \mathbf{X}) = \frac{e^{\theta_{32}}}{1 + e^{\theta_{31}} + e^{\theta_{32}} + e^{\theta_{33}}}$$
$$p_{33} = P(B_3 = 3|\mathbf{Z},\ \mathbf{X}) = P(B_3 = G|\mathbf{Z},\ \mathbf{X}) = \frac{e^{\theta_{33}}}{1 + e^{\theta_{31}} + e^{\theta_{32}} + e^{\theta_{33}}}$$
$$p_{34} = P(B_3 = 0|\mathbf{Z},\ \mathbf{X}) = P(B_3 = T|\mathbf{Z},\ \mathbf{X}) = \frac{1}{1 + e^{\theta_{31}} + e^{\theta_{32}} + e^{\theta_{33}}},$$

pois $\sum_{j=1}^{4} p_{3j} = 1$.

Observe que a estrutura de dependência presente na segunda e terceira posições do códon aparece nas probabilidades através das covariáveis Z.

Para a primeira posição do códon a probabilidade baseada na multinomial é dada por:

$\mathrm{P}(M_{11} = m_{11},\ M_{12} = m_{12},\ M_{13} = m_{13}, M_{14} = m_{14}) \propto$

$$\begin{aligned}
&\propto (p_{11})^{m_{11}}(p_{12})^{m_{12}}(p_{13})^{m_{13}}(p_{14})^{m_{14}} \\
&= \left(\frac{e^{\theta_{11}}}{1 + \sum_{j=1}^{3} e^{\theta_{1j}}}\right)^{m_{11}} \left(\frac{e^{\theta_{12}}}{1 + \sum_{j=1}^{3} e^{\theta_{1j}}}\right)^{m_{12}} \left(\frac{e^{\theta_{13}}}{1 + \sum_{j=1}^{3} e^{\theta_{1j}}}\right)^{m_{13}} \\
&\times \left(\frac{1}{1 + \sum_{j=1}^{3} e^{\theta_{1j}}}\right)^{m_{14}} \\
&= \frac{e^{\theta_{11}m_{11}} e^{\theta_{12}m_{12}} e^{\theta_{13}m_{13}}}{\left(1 + \sum_{j=1}^{3} e^{\theta_{1j}}\right)^{M_1}},
\end{aligned}$$

em que $m_{11} + m_{12} + m_{13} + m_{14} = M_1$.

Para cada códon, definimos as freqüências dos nucleotídeos em cada posição, ou seja, m_{ij} representa a freqüência da base j na posição i do códon nas seqüências de DNA. Por exemplo, m_{13} denota a freqüência do nucleotídeo G, *guanina*, na primeira posição de um determinado códon em todas as seqüências estudadas. Assim, M_i representa a soma das freqüências de todas as bases na posição i do códon.

As probabilidades para a segunda e terceira posições do códon seguem de maneira

análoga:

$$P(M_{21} = m_{21}, M_{22} = m_{22}, M_{23} = m_{23}, M_{24} = m_{24}) \propto \frac{e^{\theta_{21}m_{21}+\theta_{22}m_{22}+\theta_{23}m_{23}}}{(1+\sum_{j=1}^{3} e^{\theta_{2j}})^{M_2}} \quad e$$

$$P(M_{31} = m_{31}, M_{32} = m_{32}, M_{33} = m_{33}, M_{34} = m_{34}) \propto \frac{e^{\theta_{31}m_{31}+\theta_{32}m_{32}+\theta_{33}m_{33}}}{(1+\sum_{j=1}^{3} e^{\theta_{3j}})^{M_3}} \ .$$

As probabilidades da segunda posição do códon estão condicionadas em B_1 através dos logitos θ_{2j} e as probabilidades da terceira posição do códon estão condicionadas em B_1 e B_2 através dos logitos θ_{3j}. Ou seja, a base que aparece na segunda posição do códon depende da base que aconteceu na primeira posição do códon, e essa relação de dependência é incluída no modelo através dos logitos.

Para calcular a função de log-verossimilhança dos modelos é necessário calcular a probabilidade definida em (3.3), então, sabemos que:

$$\begin{aligned}
P(B_1, B_2, B_3|\mathbf{X}) &= P(B_1|X)P(B_2|B_1, \mathbf{X})P(B_3|B_1, B_2, \mathbf{X}) \\
&= \left[\frac{e^{\theta_{11}m_{11}+\theta_{12}m_{12}+\theta_{13}m_{13}}}{\left(1+\sum_{j=1}^{3} e^{\theta_{1j}}\right)^{M_1}}\right]\left[\frac{e^{\theta_{21}m_{21}+\theta_{22}m_{22}+\theta_{23}m_{23}}}{\left(1+\sum_{j=1}^{3} e^{\theta_{2j}}\right)^{M_2}}\right] \\
&\times \left[\frac{e^{\theta_{31}m_{31}+\theta_{32}m_{32}+\theta_{33}m_{33}}}{\left(1+\sum_{j=1}^{3} e^{\theta_{3j}}\right)^{M_3}}\right] \\
&= \frac{1}{\left(1+\sum_{j=1}^{3} e^{\theta_{1j}}\right)^{M_1}\left(1+\sum_{j=1}^{3} e^{\theta_{2j}}\right)^{M_2}\left(1+\sum_{j=1}^{3} e^{\theta_{3j}}\right)^{M_3}} \\
&\times exp\left(\sum_{i=1}^{3}\sum_{j=1}^{3}\theta_{ij}m_{ij}\right). \qquad (3.8)
\end{aligned}$$

Utilizando o resultado obtido em (3.8), podemos escrever:

$\log[\mathrm{P}(B_1,\, B_2,\, B_3|\mathbf{X})]=$

$$
\begin{aligned}
&= log[P(B_1|\mathbf{X})] + log[P(B_2|B_1,\mathbf{X})] + log[P(B_3|B_1,B_2,\mathbf{X})] \\
&= log\left[\frac{e^{\theta_{11}m_{11}+\theta_{12}m_{12}+\theta_{13}m_{13}}}{\left(1+\sum_{j=1}^{3} e^{\theta_{1j}}\right)^{M_1}}\right] + log\left[\frac{e^{\theta_{21}m_{21}+\theta_{22}m_{22}+\theta_{23}m_{23}}}{\left(1+\sum_{j=1}^{3} e^{\theta_{2j}}\right)^{M_2}}\right] \\
&+ log\left[\frac{e^{\theta_{31}.m_{31}+\theta_{32}m_{32}+\theta_{33}m_{33}}}{\left(1+\sum_{j=1}^{3} e^{\theta_{3j}}\right)^{M_3}}\right] \\
&= \theta_{11}.m_{11} + \theta_{12}m_{12} + \theta_{13}m_{13} - log\left[\left(1+\sum_{j=1}^{3} e^{\theta_{1j}}\right)^{M_1}\right] \\
&+ \theta_{21}.m_{21} + \theta_{22}m_{22} + \theta_{23}m_{23} - log\left[\left(1+\sum_{j=1}^{3} e^{\theta_{2j}}\right)^{M_2}\right] \\
&+ \theta_{31}.m_{31} + \theta_{32}m_{32} + \theta_{33}m_{33} - log\left[\left(1+\sum_{j=1}^{3} e^{\theta_{3j}}\right)^{M_3}\right] \\
&= \sum_{i=1}^{3}\sum_{j=1}^{3} \theta_{ij}.m_{ij} - M_1\, log\left[1+\sum_{j=1}^{3} e^{\theta_{1j}}\right] - M_2\, log\left[1+\sum_{j=1}^{3} e^{\theta_{2j}}\right] \\
&- M_3\, log\left[1+\sum_{j=1}^{3} e^{\theta_{3j}}\right].
\end{aligned}
\tag{3.9}
$$

Escrevendo os logitos em notação vetorial, têm-se:

$$\boldsymbol{\theta}_{(\mathbf{9\times 1})} = [\theta_{11}\ \ \theta_{12}\ \ \theta_{13}\ \ \theta_{21}\ \ \theta_{22}\ \ \theta_{23}\ \ \theta_{31}\ \ \theta_{32}\ \ \theta_{33}]'.$$

Esse vetor representa os nove logitos utilizados nos modelos, sendo três para cada posição do códon.

É fácil ver que a equação (3.9) pode ser escrita vetorialmente como:

$$\log[\mathrm{P}(B_1,\, B_2,\, B_3|\mathbf{X})]=$$

$$= \quad (m_{11}\; m_{12}\; ...\; m_{32}\; m_{33}) \begin{pmatrix} \theta_{11} \\ \theta_{12} \\ \theta_{13} \\ \vdots \\ \theta_{32} \\ \theta_{33} \end{pmatrix} - \; (M_1\; M_2\; M_3) \begin{pmatrix} log\left[1+\sum_{s=1}^{3} e^{\theta_{1s}}\right] \\ log\left[1+\sum_{s=1}^{3} e^{\theta_{2s}}\right] \\ log\left[1+\sum_{s=1}^{3} e^{\theta_{3s}}\right] \end{pmatrix}.$$

Sabe-se que existem 64 códons, no entanto, para simplificar a notação, ao invés de escrever literalmente os 64 códons: TTT, TTA, TTC, TTG,... vamos utilizar um índice l para representar os códons escritos numericamente, ou seja, códon 1, códon 2, ..., códon 64. A matriz $\boldsymbol{m}$ definida a seguir representa as freqüências das bases em cada um dos 64 códons, ou seja, k= {A, C, G e T }, i = 1, 2 e 3, l= 1, 2, ..., 64. m_{kil} representa a freqüência do nucleotídeo k, na posição i do códon l.

$$\boldsymbol{m_{(9\times 64)}} = \begin{pmatrix} m_{11,1} & m_{11,2} & m_{11,3} & \dots & m_{11(63)} & m_{11(64)} \\ m_{12,1} & m_{12,2} & m_{12,3} & \dots & m_{12(63)} & m_{12(64)} \\ m_{13,1} & m_{13,2} & m_{13,3} & \dots & m_{13(63)} & m_{13(64)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ m_{33,1} & m_{33,2} & m_{33,3} & \dots & m_{33(63)} & m_{33(64)} \end{pmatrix}$$

$$\boldsymbol{C_{(3\times 1)}} = \begin{pmatrix} log(1+e^{\theta_{11}}+e^{\theta_{12}}+e^{\theta_{13}}) \\ log(1+e^{\theta_{21}}+e^{\theta_{22}}+e^{\theta_{23}}) \\ log(1+e^{\theta_{31}}+e^{\theta_{32}}+e^{\theta_{33}}) \end{pmatrix}.$$

A matriz M indica as somas das freqüências dos nucleotídeos em cada posição do códon. Assim, $M_{i,l}$ representa a soma das freqüências dos nucleotídeos na i-ésima posição do códon l, ou seja, $M_{il} = m_{Ail} + m_{Cil} + m_{Gil} + m_{Til}$. É fácil notar que para uma única seqüência de DNA, essa é uma matriz de uns, pois sabe-se que apenas

uma das quatro freqüências é igual a um e as outras três são zero.

$$\boldsymbol{M_{(3\times 64)}} = \begin{pmatrix} M_{1,1} & M_{1,2} & M_{1,3} & \dots & M_{1(63)} & M_{1(64)} \\ M_{2,1} & M_{2,2} & M_{2,3} & \dots & M_{2(63)} & M_{2(64)} \\ M_{3,1} & M_{3,2} & M_{3,3} & \dots & M_{3(63)} & M_{3(64)} \end{pmatrix}.$$

Então, o vetor contendo as 64 log-probabilidades, ou seja, log [P(B_1, B_2, $B_3|\mathbf{X}$)] pode ser escrito matricialmente como:

$$\mathbf{log[P]}_{(64\times 1)} \;=\; \mathbf{m}'_{(64\times 9)}\boldsymbol{\theta}_{(9\times 1)} - \mathbf{M}'_{(64\times 3)}\mathbf{C}_{(3\times 1)}.$$

Então, a log-verossimilhança normalizada do modelo é dada por:

$$\begin{aligned}
\log L &= log\left[\prod_{B_1=0}^{3}\prod_{B_2=0}^{3}\prod_{B_3=0}^{3} W(B_1,B_2,B_3|\mathbf{X})^{n(B_1,B_2,B_3|\mathbf{x})}\right] \\
&= \sum_{B_1=0}^{3}\sum_{B_2=0}^{3}\sum_{B_3=0}^{3} log\left[W(B_1,B_2,B_3|\mathbf{X})^{n(B_1,B_2,B_3|\mathbf{x})}\right] \\
&= \sum_{B_1=0}^{3}\sum_{B_2=0}^{3}\sum_{B_3=0}^{3} n(B_1,B_2,B_3|\mathbf{x})\; log[W(B_1,B_2,B_3|\mathbf{X})] \\
&= \sum_{B_1=0}^{3}\sum_{B_2=0}^{3}\sum_{B_3=0}^{3} n(B_1,B_2,B_3|\mathbf{x})\; log\left[\frac{P(B_1,B_2,B_3|\mathbf{X})}{\sum_\eta P(B_1,B_2,B_3|\mathbf{X})}\right] \\
&= \sum_{B_1=0}^{3}\sum_{B_2=0}^{3}\sum_{B_3=0}^{3} n(B_1,B_2,B_3|\mathbf{x})\; log[P(B_1,B_2,B_3|\mathbf{X})] \\
&- \sum_{B_1=0}^{3}\sum_{B_2=0}^{3}\sum_{B_3=0}^{3} n(B_1,B_2,B_3|\mathbf{x}) log\left[\sum_\eta P(B_1,B_2,B_3|\mathbf{X})\right]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{B_1=0}^{3}\sum_{B_2=0}^{3}\sum_{B_3=0}^{3} n(B_1,B_2,B_3|\mathbf{x}) \sum_{i=1}^{3}\sum_{j=1}^{3}\theta_{ij}.m_{ij} - M_1\, log\left[1+\sum_{j=1}^{3}e^{\theta_{1j}}\right] \\
&- M_2\, log\left[1+\sum_{j=1}^{3}e^{\theta_{2j}}\right] - M_3\, log\left[1+\sum_{j=1}^{3}e^{\theta_{3j}}\right] - \sum_{B_1=0}^{3}\sum_{B_2=0}^{3}\sum_{B_3=0}^{3} n(B_1,B_2,B_3|\mathbf{x}) \\
&\times log\left[\sum_{\eta}\frac{exp\left(\sum_{i=1}^{3}\sum_{j=1}^{3}\theta_{ij}m_{ij}\right)}{\left(1+\sum_{j=1}^{3}e^{\theta_{1j}}\right)^{M_1}\left(1+\sum_{j=1}^{3}e^{\theta_{2j}}\right)^{M_2}\left(1+\sum_{j=1}^{3}e^{\theta_{3j}}\right)^{M_3}}\right]. \qquad (3.10)
\end{aligned}
$$

em que η é o conjunto dos códons efetivos.

A log-verossimilhança sem a normalização é dada por:

$$
\begin{aligned}
\log L &= \sum_{B_1=0}^{3}\sum_{B_2=0}^{3}\sum_{B_3=0}^{3} n(B_1,B_2,B_3|\mathbf{x}) \sum_{i=1}^{3}\sum_{j=1}^{3}\theta_{ij}.m_{ij} - M_1\, log\left[1+\sum_{j=1}^{3}e^{\theta_{1j}}\right] \\
&- M_2\, log\left[1+\sum_{j=1}^{3}e^{\theta_{2j}}\right] - M_3\, log\left[1+\sum_{j=1}^{3}e^{\theta_{3j}}\right]. \qquad (3.11)
\end{aligned}
$$

Para escrever matricialmente a log-verossimilhança é necessário definir um vetor $\boldsymbol{n}$ com as freqüências dos 64 códons, dados por,

$$\mathbf{n} = [n_1 \;\; n_2 \;\; ... \;\; n_{64}]'.$$

Além disso, é necessário definir uma matriz $\mathbf{S}$ que nada mais é do que a repetição de 64 linhas de uma expressão:

$$
\mathbf{S} = \begin{pmatrix} log\sum_{\eta} P(B_1,B_2,B_3|\mathbf{X}) \\ log\sum_{\eta} P(B_1,B_2,B_3|\mathbf{X}) \\ log\sum_{\eta} P(B_1,B_2,B_3|\mathbf{X}) \\ \vdots \\ log\sum_{\eta} P(B_1,B_2,B_3|\mathbf{X}) \end{pmatrix} = \begin{pmatrix} log\sum_{\eta}[exp\,(\mathbf{m}'\boldsymbol{\theta} - \mathbf{M}'\mathbf{C})] \\ log\sum_{\eta}[exp\,(\mathbf{m}'\boldsymbol{\theta} - \mathbf{M}'\mathbf{C})] \\ log\sum_{\eta}[exp\,(\mathbf{m}'\boldsymbol{\theta} - \mathbf{M}'\mathbf{C})] \\ \vdots \\ log\sum_{\eta}[exp\,(\mathbf{m}'\boldsymbol{\theta} - \mathbf{M}'\mathbf{C})] \end{pmatrix}.
$$

Assim a log-verossimilhança (3.10) escrita de forma matricial é dada pela seguinte

fórmula:

$$\begin{aligned} log\ L(\boldsymbol{\theta}) &= \mathbf{n}'.[\mathbf{m}'\boldsymbol{\theta} - \mathbf{M}'\mathbf{C}] - \mathbf{n}.\mathbf{S} \\ &= \mathbf{n}'.[\mathbf{m}'\boldsymbol{\theta} - \mathbf{M}'\mathbf{C}] - \mathbf{n}'.[log(\mathbf{11_{60}}'(\mathbf{m}'\boldsymbol{\theta} - \mathbf{M}'\mathbf{C}))\mathbf{11_{64}}], \end{aligned}$$

em que 11_{60} é um vetor de uns com dimensão 60×1 e 11_{64} é um vetor de uns com dimensão 64×1.

Foram desenvolvidos programas em MatLab que calculam as log-verossimilhanças e as estimativas foram obtidas através de uma rotina de maximização do MatLab. Os resultados dos modelos aplicados à dados reais podem ser vistos no capítulo 4.

Dentre os modelos a serem analisados, o modelo mais simples é o que não considera nenhuma estrutura de dependência, ou seja, o modelo independente. Por esse motivo, ele será utilizado para ilustrar os cálculos das derivadas necessárias para se obter os estimadores de máxima verossimilhança dos parâmetros.

Assim, considerando o modelo independente sem covariáveis e derivando-se o log da probabilidade do códon (3.9) com relação a α_{ij}, obtém-se:

$$\frac{\partial logP(B_1, B_2, B_3|X)}{\partial \alpha_{ij}} = \sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij} \times \left[\frac{\partial \theta_{ij}}{\partial \alpha_{ij}}\right] - \frac{M_i\ exp(\theta_{ij})}{(1 + e^{\theta_{i1}} + e^{\theta_{i2}} + e^{\theta_{i3}})} \times \left[\frac{\partial \theta_{ij}}{\partial \alpha_{ij}}\right].$$

Sabe-se que $\frac{\partial \theta_{ij}}{\partial \alpha_{ij}} = 1$ para qualquer modelo, então, a derivada é dada por:

$$\frac{\partial logP(B_1, B_2, B_3|X)}{\partial \alpha_{ij}} = \sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij} - \frac{M_i\ exp(\theta_{ij})}{(1 + e^{\theta_{i1}} + e^{\theta_{i2}} + e^{\theta_{i3}})}. \qquad (3.12)$$

Considerando o modelo independente sem covariáveis e derivando-se a função de log-verossimilhança (3.10) com relação a α_{ij}, obtém-se:

$$\begin{aligned} \frac{\partial logL}{\partial \alpha_{ij}} &= \sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij} - \frac{M_i\ exp(\alpha_{ij})}{(1 + e^{\theta_{i1}} + e^{\theta_{i2}} + e^{\theta_{i3}})} \\ &- \frac{1}{\sum_\eta P(B_1, B_2, B_3|X)} \times \sum_\eta \frac{\partial P(B_1, B_2, B_3|X)}{\partial \alpha_{ij}}. \end{aligned}$$

Sabe-se que:

$$P(B_1, B_2, B_3|X) \;=\; exp\,(log\,[P(B_1, B_2, B_3|X)])\;\; e$$

$$\begin{aligned}
\frac{\partial P(B_1, B_2, B_3|X)}{\partial \alpha_{ij}} &= \frac{\partial\;\; [exp\,(log\,[P(B_1, B_2, B_3|X)])]}{\partial \alpha_{ij}} \\
&= exp\,(log\,[P(B_1, B_2, B_3|X)]) \times \frac{\partial\; log[P(B_1, B_2, B_3|X)]}{\partial \alpha_{ij}} \\
&= P(B_1, B_2, B_3|X)) \times (3.12).
\end{aligned}$$

Então, a derivada é dada por:

$$\begin{aligned}
\frac{\partial logL}{\partial \alpha_{ij}} &= \sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij} - \frac{M_i\; exp\,(\alpha_{ij})}{(1 + e^{\alpha_{i1}} + e^{\alpha_{i2}} + e^{\alpha_{i3}})} \\
&- \frac{1}{\sum_\eta P(B_1, B_2, B_3|X)} \times \sum_\eta P(B_1, B_2, B_3|X) \times (3.12) \\
&= \sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij} - \frac{M_i\; exp\,(\alpha_{ij})}{(1 + e^{\alpha_{i1}} + e^{\alpha_{i2}} + e^{\alpha_{i3}})} - \frac{1}{\sum_\eta P(B_1, B_2, B_3|X)} \\
&\times \sum_\eta \left[P(B_1, B_2, B_3|X) \sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij} - \frac{M_i\; exp\,(\alpha_{ij})}{(1 + e^{\alpha_{i1}} + e^{\alpha_{i2}} + e^{\alpha_{i3}})} \right] (3.13)
\end{aligned}$$

Para obter o estimador de máxima verossimilhança de α_{ij}, basta igualar a equação (3.13) a zero e isolar α_{ij}. No entanto, é fácil notar que o estimador de máxima verossimilhança de α_{ij} não posui forma fechada, ou seja, o estimador do parâmetro depende do próprio parâmetro. Por esse motivo é necessário utilizar métodos iterativos para achar as estimativas dos parâmetros.

Se considerarmos agora o modelo independente com covariáveis, seria interes-

sante achar os estimadores dos β's, assim, derivando (3.9) com relação a β_k, têm-se:

$$\begin{aligned}\frac{\partial logP(B_1,B_2,B_3|X)}{\partial\beta_k} &= \sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij}\times\left[\frac{\partial\theta_{ij}}{\partial\beta_k}\right] - X_k\sum_{i=1}^{3} M_i\left(\frac{e^{\theta_{i1}}+e^{\theta_{i2}}+e^{\theta_{i3}}}{1+e^{\theta_{i1}}+e^{\theta_{i2}}+e^{\theta_{i3}}}\right)\\ &= X_k\sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij} - X_k\sum_{i=1}^{3} M_i\left(\frac{e^{\theta_{i1}}+e^{\theta_{i2}}+e^{\theta_{i3}}}{1+e^{\theta_{i1}}+e^{\theta_{i2}}+e^{\theta_{i3}}}\right) \quad (3.14)\end{aligned}$$

Agora, derivando-se a log-verossimilhança (3.10) com relação a β_k, temos:

$$\begin{aligned}\frac{\partial logL}{\partial\beta_k} &= X_k\sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij} - X_k\sum_{i=1}^{3} M_i\left(\frac{e^{\theta_{i1}}+e^{\theta_{i2}}+e^{\theta_{i3}}}{1+e^{\theta_{i1}}+e^{\theta_{i2}}+e^{\theta_{i3}}}\right)\\ &- \frac{\sum_\eta \frac{\partial P(B_1,B_2,B_3|X)}{\partial\beta_k}}{\sum_\eta P(B_1,B_2,B_3|X)}\\ &= X_k\sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij} - X_k\sum_{i=1}^{3} M_i\left(\frac{e^{\theta_{i1}}+e^{\theta_{i2}}+e^{\theta_{i3}}}{1+e^{\theta_{i1}}+e^{\theta_{i2}}+e^{\theta_{i3}}}\right)\\ &- \frac{\sum_\eta P(B_1,B_2,B_3|X)\times(3.14)}{\sum_\eta P(B_1,B_2,B_3|X)}\\ &= X_k\sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij} - X_k\sum_{i=1}^{3} M_i\left(\frac{e^{\theta_{i1}}+e^{\theta_{i2}}+e^{\theta_{i3}}}{1+e^{\theta_{i1}}+e^{\theta_{i2}}+e^{\theta_{i3}}}\right)\\ &- \frac{1}{\sum_\eta P(B_1,B_2,B_3|X)}\times\sum_\eta\left[P(B_1,B_2,B_3|X)\right.\\ &\times \left.\left(X_k\sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij} - X_k\sum_{i=1}^{3} M_i\left(\frac{e^{\theta_{i1}}+e^{\theta_{i2}}+e^{\theta_{i3}}}{1+e^{\theta_{i1}}+e^{\theta_{i2}}+e^{\theta_{i3}}}\right)\right)\right].\end{aligned}$$

Além dos estimadores de máxima verossimilhança dos parâmetros, seria interessante também obter a matriz de informação de Fisher para calcular os erros padrão das estimativas dos parâmetros. No entanto, alguns cálculos das segundas derivadas e derivadas cruzadas são muito extensos e os cálculos dos γ's são diferentes para cada um dos modelos, por isso a notação matricial será introduzida para simplificar os cálculos e apresentar os resultados de maneira mais elegante. Para ilustrar o cálculo dos estimadores dos γ's vamos utilizar a estrutura de dependência do modelo de Markov e, de agora em diante, utilizaremos a função de log-verossimilhança sem a

normalização (3.11).

O vetor de interceptos é dado por

$$\boldsymbol{\alpha}_{(9\times 1)} = (\alpha_{11}\ \alpha_{12}\ \alpha_{13}\ \alpha_{21}\ \alpha_{22}\ \alpha_{23}\ \alpha_{31}\ \alpha_{32}\ \alpha_{33})'.$$

Assim, têm-se,

$$\frac{\partial\,\boldsymbol{\theta}}{\partial\,\boldsymbol{\alpha}'} = \begin{pmatrix} \frac{\partial\theta_{11}}{\partial\alpha_{11}} & \frac{\partial\theta_{11}}{\partial\alpha_{12}} & \cdots & \frac{\partial\theta_{11}}{\partial\alpha_{33}} \\ \frac{\partial\theta_{12}}{\partial\alpha_{11}} & \frac{\partial\theta_{12}}{\partial\alpha_{12}} & \cdots & \frac{\partial\theta_{12}}{\partial\alpha_{33}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial\theta_{33}}{\partial\alpha_{11}} & \frac{\partial\theta_{33}}{\partial\alpha_{12}} & \cdots & \frac{\partial\theta_{33}}{\partial\alpha_{33}} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = Diag(1).$$

O vetor de β's é definido por

$$\boldsymbol{\beta}_{(p\times 1)} = (\beta_1\ \beta_2\ \ldots\ \beta_p)'.$$

As derivadas com relação aos β's são dadas por

$$\frac{\partial\,\boldsymbol{\theta}}{\partial\,\boldsymbol{\beta}'} = \begin{pmatrix} \frac{\partial\theta_{11}}{\partial\beta_1} & \frac{\partial\theta_{11}}{\partial\beta_2} & \cdots & \frac{\partial\theta_{11}}{\partial\beta_p} \\ \frac{\partial\theta_{12}}{\partial\beta_1} & \frac{\partial\theta_{12}}{\partial\beta_2} & \cdots & \frac{\partial\theta_{12}}{\partial\beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial\theta_{33}}{\partial\beta_1} & \frac{\partial\theta_{33}}{\partial\beta_2} & \cdots & \frac{\partial\theta_{33}}{\partial\beta_p} \end{pmatrix} = \begin{pmatrix} X_1 & X_2 & \cdots & X_p \\ X_1 & X_2 & \cdots & X_p \\ \vdots & \vdots & \ddots & \vdots \\ X_1 & X_2 & \cdots & X_p \end{pmatrix} = \mathbf{X}.$$

O vetor de γ's é dado por

$$\boldsymbol{\gamma}_{(6\times 1)} = [\gamma_{21}\ \gamma_{22}\ \gamma_{23}\ \gamma_{31}\ \gamma_{32}\ \gamma_{33}]'.$$

Os logitos para o modelo de Markov são dados por

$$\begin{aligned}
\theta_{11} &= \alpha_{11} + \beta_1 X_1 + ... + \beta_p X_p \\
\theta_{12} &= \alpha_{12} + \beta_1 X_1 + ... + \beta_p X_p \\
\theta_{13} &= \alpha_{13} + \beta_1 X_1 + ... + \beta_p X_p \\
\theta_{21} &= \alpha_{21} + \gamma_{21} Z_{11} + \gamma_{22} Z_{12} + \gamma_{23} Z_{13} + \beta_1 X_1 + ... + \beta_p X_p \\
\theta_{22} &= \alpha_{22} + \gamma_{21} Z_{11} + \gamma_{22} Z_{12} + \gamma_{23} Z_{13} + \beta_1 X_1 + ... + \beta_p X_p \\
\theta_{23} &= \alpha_{23} + \gamma_{21} Z_{11} + \gamma_{22} Z_{12} + \gamma_{23} Z_{13} + \beta_1 X_1 + ... + \beta_p X_p \\
\theta_{31} &= \alpha_{31} + \gamma_{31} Z_{21} + \gamma_{32} Z_{22} + \gamma_{33} Z_{23} + \beta_1 X_1 + ... + \beta_p X_p \\
\theta_{32} &= \alpha_{32} + \gamma_{31} Z_{21} + \gamma_{32} Z_{22} + \gamma_{33} Z_{23} + \beta_1 X_1 + ... + \beta_p X_p \\
\theta_{33} &= \alpha_{33} + \gamma_{31} Z_{21} + \gamma_{32} Z_{22} + \gamma_{33} Z_{23} + \beta_1 X_1 + ... + \beta_p X_p. \qquad (3.15)
\end{aligned}$$

Assim, é fácil ver que os logitos podem ser escritos de forma matricial da seguinte maneira,

$$\boldsymbol{\theta}_{(9\times 1)} = \boldsymbol{\alpha}_{(9\times 1)} + \mathbf{Z}_{M(9\times 6)}.\boldsymbol{\gamma}_{(6\times 1)} + \boldsymbol{\beta}_{(9\times p)}.\mathbf{X}_{(p\times 1)}.$$

Para cada modelo com estrutura de dependência teremos uma matriz $\mathbf{Z}$ diferente.

Então, as derivadas parciais são dadas por

$$\frac{\partial\, \boldsymbol{\theta}}{\partial\, \boldsymbol{\alpha}'} = \mathbf{11}; \qquad \frac{\partial\, \boldsymbol{\theta}}{\partial\, \boldsymbol{\beta}'} = \mathbf{X}; \qquad \frac{\partial\, \boldsymbol{\theta}}{\partial\, \boldsymbol{\gamma}'} = \mathbf{Z}.$$

Então, a matriz de derivadas parciais, $\boldsymbol{Z_M}$ para o modelo de Markov é dada por

$$\frac{\partial\,\boldsymbol{\theta}}{\partial\,\boldsymbol{\gamma'}} = \begin{pmatrix} \frac{\partial\theta_{11}}{\partial\gamma_{21}} & \frac{\partial\theta_{11}}{\partial\gamma_{22}} & \cdots & \frac{\partial\theta_{11}}{\partial\gamma_{33}} \\ \frac{\partial\theta_{12}}{\partial\gamma_{21}} & \frac{\partial\theta_{12}}{\partial\gamma_{22}} & \cdots & \frac{\partial\theta_{12}}{\partial\gamma_{33}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial\theta_{33}}{\partial\gamma_{21}} & \frac{\partial\theta_{33}}{\partial\gamma_{22}} & \cdots & \frac{\partial\theta_{33}}{\partial\gamma_{33}} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ Z_{11} & Z_{12} & Z_{13} & 0 & 0 & 0 \\ Z_{11} & Z_{12} & Z_{13} & 0 & 0 & 0 \\ Z_{11} & Z_{12} & Z_{13} & 0 & 0 & 0 \\ 0 & 0 & 0 & Z_{21} & Z_{22} & Z_{23} \\ 0 & 0 & 0 & Z_{21} & Z_{22} & Z_{23} \\ 0 & 0 & 0 & Z_{21} & Z_{22} & Z_{23} \end{pmatrix} = \boldsymbol{Z_M}.$$

As derivadas de C com relação a α são dadas por,

$$\frac{\partial\,\mathbf{C}}{\partial\,\boldsymbol{\alpha}} = \begin{pmatrix} \frac{\partial log(1+e^{\theta_{11}}+e^{\theta_{12}}+e^{\theta_{13}})}{\partial\alpha_{11}} & \frac{\partial log(1+e^{\theta_{11}}+e^{\theta_{12}}+e^{\theta_{13}})}{\partial\alpha_{12}} & \cdots & \frac{\partial log(1+e^{\theta_{11}}+e^{\theta_{12}}+e^{\theta_{13}})}{\partial\alpha_{33}} \\ \frac{\partial log(1+e^{\theta_{21}}+e^{\theta_{22}}+e^{\theta_{23}})}{\partial\alpha_{11}} & \frac{\partial log(1+e^{\theta_{21}}+e^{\theta_{22}}+e^{\theta_{23}})}{\partial\alpha_{12}} & \cdots & \frac{\partial log(1+e^{\theta_{21}}+e^{\theta_{22}}+e^{\theta_{23}})}{\partial\alpha_{33}} \\ \frac{\partial log(1+e^{\theta_{31}}+e^{\theta_{32}}+e^{\theta_{33}})}{\partial\alpha_{11}} & \frac{\partial log(1+e^{\theta_{31}}+e^{\theta_{32}}+e^{\theta_{33}})}{\partial\alpha_{12}} & \cdots & \frac{\partial log(1+e^{\theta_{31}}+e^{\theta_{32}}+e^{\theta_{33}})}{\partial\alpha_{33}} \end{pmatrix}.$$

Seja $t_i = e^{\theta_{i1}} + e^{\theta_{i2}} + e^{\theta_{i3}}$, $i = 1$, 2 e 3.

A matriz de derivadas é bloco diagonal, com diagonais determinadas pelos vetores D_1, D_2 e D_3, ou seja,

$$D_i = \begin{bmatrix} \frac{e^{\theta_{i1}}}{(1+t_i)} & \frac{e^{\theta_{i2}}}{(1+t_i)} & \frac{e^{\theta_{i3}}}{(1+t_i)} \end{bmatrix} \; e$$

$$\frac{\partial\,\mathbf{C}}{\partial\,\boldsymbol{\alpha}} = \begin{pmatrix} \underline{D}_1 & \underline{0} & \underline{0} \\ \underline{0} & \underline{D}_2 & \underline{0} \\ \underline{0} & \underline{0} & \underline{D}_3 \end{pmatrix} = \mathbf{D}.$$

É fácil notar que a matriz de segundas derivadas parciais também é bloco dia-

gonal, com diagonais determinadas por H_i, $i = 1$, 2 e 3.

$$H_i \;=\; \left[\frac{e^{\theta_{i1}}}{(1+t_i)^2} \quad \frac{e^{\theta_{i2}}}{(1+t_i)^2} \quad \frac{e^{\theta_{i3}}}{(1+t_i)^2}\right] \;\; e$$

$$\frac{\partial^2 \mathbf{C}}{\partial \boldsymbol{\alpha}\, \partial \boldsymbol{\alpha}^{\boldsymbol{T}}} \;=\; \begin{pmatrix} \underline{H}_1 & \underline{0} & \underline{0} \\ \underline{0} & \underline{H}_2 & \underline{0} \\ \underline{0} & \underline{0} & \underline{H}_3 \end{pmatrix} = \mathbf{H}.$$

A matriz de derivadas parciais de C com relação aos β's é dada por

$$\frac{\partial\, \mathbf{C}}{\partial\, \boldsymbol{\beta}} \;=\; \begin{pmatrix} \frac{X_1\, t_1}{1+t_1} & \cdots & \frac{X_p\, t_1}{1+t_1} \\ \frac{X_1\, t_2}{1+t_2} & \cdots & \frac{X_p\, t_2}{1+t_2} \\ \frac{X_1\, t_3}{1+t_3} & \cdots & \frac{X_p\, t_3}{1+t_3} \end{pmatrix} = \mathbf{Q}.$$

A matriz de segundas derivadas parciais é dada por

$$\frac{\partial^2\, \mathbf{C}}{\partial\, \boldsymbol{\beta}\, \partial\, \boldsymbol{\beta}^{\boldsymbol{T}}} \;=\; \begin{pmatrix} \frac{-X_1^2\, t_1}{(1+t_1)^2} & \cdots & \frac{-X_p^2\, t_1}{(1+t_1)^2} \\ \frac{-X_1^2\, t_2}{(1+t_2)^2} & \cdots & \frac{-X_p^2\, t_2}{(1+t_2)^2} \\ \frac{-X_1^2\, t_3}{(1+t_3)^2} & \cdots & \frac{-X_p^2\, t_3}{(1+t_3)^2} \end{pmatrix} = \mathbf{R}.$$

A matriz de derivadas de $\mathbf{C}$ com relação aos γ's para o modelo de Markov é dada por

$$\frac{\partial\, \mathbf{C}}{\partial\, \boldsymbol{\gamma}} \;=\; \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{Z_{11}\, t_2}{1+t_2} & \frac{Z_{12}\, t_2}{1+t_2} & \frac{Z_{13}\, t_2}{1+t_2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{Z_{21}\, t_3}{1+t_3} & \frac{Z_{22}\, t_3}{1+t_3} & \frac{Z_{23}\, t_3}{1+t_3} \end{pmatrix} = \mathbf{U}.$$

A matriz de segundas derivadas é dada por

$$\frac{\partial^2\, \mathbf{C}}{\partial\, \boldsymbol{\gamma}\, \partial\, \boldsymbol{\gamma}^{\boldsymbol{T}}} \;=\; \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{-Z_{11}^2\, t_2}{(1+t_2)^2} & \frac{-Z_{12}^2\, t_2}{(1+t_2)^2} & \frac{-Z_{13}^2\, t_2}{(1+t_2)^2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{-Z_{21}^2\, t_3}{(1+t_3)^2} & \frac{-Z_{22}^2\, t_3}{(1+t_3)^2} & \frac{-Z_{23}^2\, t_3}{(1+t_3)^2} \end{pmatrix} = \mathbf{W}.$$

Para obter a matriz de Informação de Fisher será necessário derivar a log verossimilhança com relação à todos os parâmetros:

$$\begin{aligned}\frac{\partial\, logL}{\partial\, \boldsymbol{\alpha}} &= \mathbf{n}'\left(\mathbf{m}'\left[\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\alpha}}\right] - \mathbf{M}'\left[\frac{\partial\, \mathbf{C}}{\partial \boldsymbol{\alpha}}\right]\right) \\ &= \mathbf{n}'(\mathbf{m}'\mathbf{11} - \mathbf{M}'\mathbf{D}).\end{aligned}$$

$$\frac{\partial^2\, logL}{\partial\, \boldsymbol{\alpha}\;\partial\, \boldsymbol{\alpha}^{\boldsymbol{T}}} = \mathbf{n}'\left(-\mathbf{M}'\left[\frac{\partial\, \mathbf{D}}{\partial \boldsymbol{\alpha}}\right]\right) = -\mathbf{n}'(\mathbf{M}'\mathbf{H}). \tag{3.16}$$

$$\begin{aligned}\frac{\partial\, logL}{\partial\, \boldsymbol{\beta}} &= \mathbf{n}'\left(\mathbf{m}'\left[\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}}\right] - M'\left[\frac{\partial\, \mathbf{C}}{\partial \boldsymbol{\beta}}\right]\right) \\ &= \mathbf{n}'(\mathbf{m}'\mathbf{X} - \mathbf{M}'\mathbf{Q}).\end{aligned}$$

$$\frac{\partial^2\, logL}{\partial \boldsymbol{\beta}\;\partial \boldsymbol{\beta}^{\boldsymbol{T}}} = \mathbf{n}'\left(-\mathbf{M}'\left[\frac{\partial\, \mathbf{Q}}{\partial \boldsymbol{\beta}}\right]\right) = -\mathbf{n}'(\mathbf{M}'\mathbf{R}). \tag{3.17}$$

$$\begin{aligned}\frac{\partial\, logL}{\partial\, \boldsymbol{\gamma}} &= \mathbf{n}'\left(\mathbf{m}'\left[\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\gamma}}\right] - M'\left[\frac{\partial\, \mathbf{C}}{\partial \boldsymbol{\gamma}}\right]\right) \\ &= \mathbf{n}'(\mathbf{m}'\mathbf{Z} - \mathbf{M}'\mathbf{U}).\end{aligned}$$

$$\frac{\partial^2\, logL}{\partial \boldsymbol{\gamma}\;\partial \boldsymbol{\gamma}^{\boldsymbol{T}}} = \mathbf{n}'\left(-\mathbf{M}'\left[\frac{\partial\, \mathbf{U}}{\partial \boldsymbol{\gamma}}\right]\right) = -\mathbf{n}'(\mathbf{M}'\mathbf{W}). \tag{3.18}$$

As derivadas parciais cruzadas são dadas por

$$\frac{\partial \mathbf{D}}{\partial \boldsymbol{\beta}} = \begin{pmatrix} E_1 & \underline{0} & \underline{0} \\ \underline{0} & E_2 & \underline{0} \\ \underline{0} & \underline{0} & E_3 \end{pmatrix} = \mathbf{V},$$

em que E_i, $i = 1, 2, 3$ é dada por

$$\boldsymbol{E_i} = \begin{pmatrix} \frac{X_1 e^{\theta_{i1}}}{(1+t_i)^2} & \frac{X_1 e^{\theta_{i2}}}{(1+t_i)^2} & \frac{X_1 e^{\theta_{i3}}}{(1+t_i)^2} \\ \vdots & \vdots & \vdots \\ \frac{X_p\, e^{\theta_{i1}}}{(1+t_i)^2} & \frac{X_p\, e^{\theta_{i2}}}{(1+t_i)^2} & \frac{X_p\, e^{\theta_{i3}}}{(1+t_i)^2} \end{pmatrix}$$

e $\underline{\mathbf{0}}$ é um matriz de zeros com dimensão p $\times$ 3. Derivando-se $\mathbf{D}$ com relação a γ, obtemos

$$\frac{\partial \mathbf{D}}{\partial \boldsymbol{\gamma}} = \begin{pmatrix} \underline{0} & \underline{0} & \underline{0} \\ \underline{0} & \underline{0} & \underline{0} \\ \underline{0} & \underline{A}_2 & \underline{0} \\ \underline{0} & \underline{0} & \underline{0} \\ \underline{0} & \underline{0} & \underline{0} \\ \underline{0} & \underline{0} & \underline{A}_3 \end{pmatrix} = \mathbf{G},$$

em que A_i, $i = 2, 3$ é definida como

$$\boldsymbol{A_i} = \begin{pmatrix} \frac{Z_{(i-1)1} e^{\theta_{i1}}}{(1+t_i)^2} & \frac{Z_{(i-1)1} e^{\theta_{i2}}}{(1+t_i)^2} & \frac{Z_{(i-1)1} e^{\theta_{i3}}}{(1+t_i)^2} \\ \frac{Z_{(i-1)2} e^{\theta_{i1}}}{(1+t_i)^2} & \frac{Z_{(i-1)2} e^{\theta_{i2}}}{(1+t_i)^2} & \frac{Z_{(i-1)2} e^{\theta_{i3}}}{(1+t_i)^2} \\ \frac{Z_{(i-1)3} e^{\theta_{i1}}}{(1+t_i)^2} & \frac{Z_{(i-1)3} e^{\theta_{i2}}}{(1+t_i)^2} & \frac{Z_{(i-1)3} e^{\theta_{i3}}}{(1+t_i)^2} \end{pmatrix}$$

e $\underline{\mathbf{0}}$ agora é uma matriz de zeros com dimensão 3 $\times$ 3.

$$\frac{\partial \mathbf{U}}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \underline{0} & \underline{0} \\ \underline{B}_2 & \underline{0} \\ \underline{0} & \underline{B}_3 \end{pmatrix} = \mathbf{T}$$

e B_i, $i = 2, 3$ é definida como

$$\boldsymbol{B_i} = \begin{pmatrix} \frac{Z_{(i-1)1} X_1 t_i}{(1+t_i)^2} & \frac{Z_{(i-1)2} X_1 t_i}{(1+t_i)^2} & \frac{Z_{(i-1)3} X_1 t_i}{(1+t_i)^2} \\ \frac{Z_{(i-1)1} X_2 t_i}{(1+t_i)^2} & \frac{Z_{(i-1)2} X_2 t_i}{(1+t_i)^2} & \frac{Z_{(i-1)3} X_2 t_i}{(1+t_i)^2} \\ \vdots & \vdots & \vdots \\ \frac{Z_{(i-1)1} X_p t_i}{(1+t_i)^2} & \frac{Z_{(i-1)2} X_p t_i}{(1+t_i)^2} & \frac{Z_{(i-1)3} X_3 t_i}{(1+t_i)^2} \end{pmatrix}$$

onde agora $\mathbf{O}$ representa uma matriz de zeros com dimensão p $\times$ 3.

$$\begin{aligned}\frac{\partial^2 logL}{\partial \boldsymbol{\alpha}\, \partial \boldsymbol{\beta}^{\boldsymbol{T}}} &= \frac{\partial^2 logL}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\alpha}^{\boldsymbol{T}}} = \frac{\partial}{\partial \boldsymbol{\beta}}(\mathbf{n}'(\mathbf{m}'\mathbf{11} - \mathbf{M}'\mathbf{D})) \\ &= -\mathbf{n}'\left(\mathbf{M}'\left[\frac{\partial \mathbf{D}}{\partial \boldsymbol{\beta}}\right]\right) = -\mathbf{n}(\mathbf{M}'\mathbf{V}). \end{aligned} \tag{3.19}$$

$$\begin{aligned}\frac{\partial^2 logL}{\partial \boldsymbol{\gamma}\, \partial \boldsymbol{\alpha}^{\boldsymbol{T}}} &= \frac{\partial}{\partial \boldsymbol{\gamma}}\left(n'\left(\mathbf{m}'\mathbf{11} - \mathbf{M}'\mathbf{D}\right)\right) \\ &= \frac{\partial}{\partial \boldsymbol{\gamma}}(\mathbf{n}'(-\mathbf{M}'\mathbf{D})) = -\mathbf{n}'\left(\mathbf{M}'\left[\frac{\partial \mathbf{D}}{\partial \boldsymbol{\gamma}}\right]\right) = -\mathbf{n}'(\mathbf{M}'\mathbf{G}). \end{aligned} \tag{3.20}$$

$$\begin{aligned}\frac{\partial^2 logL}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\gamma}^{\boldsymbol{T}}} &= \frac{\partial}{\partial \boldsymbol{\beta}}\left(n'(m'\mathbf{Z} - \mathbf{M}'\mathbf{U})\right) = \frac{\partial}{\partial \boldsymbol{\beta}}(\mathbf{n}'(-\mathbf{M}'\mathbf{U})) \\ &= -\mathbf{n}'\left(\mathbf{M}'\left[\frac{\partial \mathbf{U}}{\partial \boldsymbol{\beta}}\right]\right) = -\mathbf{n}'(\mathbf{M}'\mathbf{T}). \end{aligned} \tag{3.21}$$

Utilizando as informações obtidas em (3.16), (3.17), (3.18), (3.19), (3.20) e (3.21) a matriz de Informação de Fisher é dada por:

$$\boldsymbol{I_F} = -\begin{pmatrix} \frac{\partial^2 logL}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^{\boldsymbol{T}}} & \frac{\partial^2 logL}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\beta}^{\boldsymbol{T}}} & \frac{\partial^2 logL}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\gamma}^{\boldsymbol{T}}} \\ \frac{\partial^2 logL}{\partial \boldsymbol{\beta}\partial \boldsymbol{\alpha}^{\boldsymbol{T}}} & \frac{\partial^2 logL}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\boldsymbol{T}}} & \frac{\partial^2 logL}{\partial \boldsymbol{\beta}\partial \boldsymbol{\gamma}^{\boldsymbol{T}}} \\ \frac{\partial^2 logL}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\alpha}^{\boldsymbol{T}}} & \frac{\partial^2 logL}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\beta}^{\boldsymbol{T}}} & \frac{\partial^2 logL}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}^{\boldsymbol{T}}} \end{pmatrix} = \begin{pmatrix} \mathbf{n}'(\mathbf{M}'\mathbf{H}) & \mathbf{n}'(\mathbf{M}'\mathbf{V}) & n'(\mathbf{M}'\mathbf{G}) \\ \mathbf{n}'(\mathbf{M}'\mathbf{V}) & \mathbf{n}'(\mathbf{M}'\mathbf{R}) & n'(\mathbf{M}'\mathbf{T}) \\ \mathbf{n}'(\mathbf{M}'\mathbf{G}) & \mathbf{n}'(\mathbf{M}'\mathbf{T}) & n'(\mathbf{M}'\mathbf{W}) \end{pmatrix}.$$

A matriz assintótica de covariâncias dos estimadores de máxima verossimilhança é o inverso da matriz de informação, $\boldsymbol{\Sigma}(\boldsymbol{\lambda}) = (\boldsymbol{I_F})^{-1}$.

4 Aplicação

Para ilustrar o uso do modelo logístico regressivo em biologia molecular e comparar os modelos estudados nos capítulos anteriores, foi feita uma análise das freqüências dos códons do gene **NADH4** (dehidrogenase subunidade 4) do genoma mitocondrial humano, inicialmente apenas para a *seqüência de referência de cambridge (SRC)*. O gene **NADH4** começa no nucleotídeo de posição 10.760 e termina na posição 12.137 do genoma mitocondrial humano. Assim, esse gene possui 1.377 pares de bases e codifica 459 aminoácidos. Esse gene foi escolhido para a aplicação, pois ele tem um número razoável de pares de bases e por possuir uma doença causada por mutação ligada à ele.

A SRC para DNA mitocondrial humano foi primeiramente publicada em 1981 como um precursor do projeto genoma humano. Um grupo sob a supervisão do Dr. Fred Sanger na Universidade de Cambridge sequenciou o genoma mitocondrial de um indivíduo de descendência européia durante a década de 1970, determinando que a seqüência tinha 16.568 pares de bases, contendo 37 genes. Posteriormente a seqüência foi revisada e publicada por Richard Andrews em 1999. A SRC continua a ser indispensável para estudos de evolução humana, genética populacional e doenças mitocondriais. Foi reconhecido, no entanto, que a SRC difere em vários nucleotídeos se comparada com outras seqüências de DNA mitocondrial (Andrews et al., 1999). A SRC de Anderson et al.(1981) é aceita mundialmente como referência de seqüência de DNA mitocondrial (DNAmit) humano.

Algumas comparações entre várias seqüências de DNA do gene NADH4 com a SRC foram feitas para analisar de maneira exploratória as freqüências dos códons.

Depois da análise exploratória, foram ajustados oito modelos para analisar as freqüências dos códons da seqüência do DNA mitocondrial da referência de Cambridge. Após essa análise outras seqüências de DNA foram incluídas na amostra e os modelos foram ajustados novamente. Todos os modelos estudados estão aninhados com o primeiro modelo, o modelo aditivo, que é o mais geral de todos. Em nenhum dos modelos foi considerado as interações. Para a análise dos dados foram desenvolvidos programas em MatLab que calculam as log-verossimilhanças e posteriormente, utilizando uma rotina de maximização obtém-se as estimativas dos parâmetros. O programa em MatLab para o modelo aditivo pode ser visto no apêndice.

Depois dos ajustes dos modelos foram feitas análises com testes da razão de verossimilhanças e o critério de Akaike para averiguar qual o modelo mais parcimonioso. Além disso, algumas medidas de diagnóstico foram calculadas para averiguar qual é o modelo que melhor se ajusta aos dados.

4.1 Neuropatia Ótica Hereditária de Leber

Mutações no DNAmit podem induzir doenças mitocondriais com herança materna que se expressam durante diferentes fases da vida e que envolvem predominantemente os músculos e o sistema nervoso, por serem os tecidos que mais precisam de energia.

LHON (*Leber's Hereditary Optic Neuropathy*) é uma doença de herança materna que provaca a degeneração do nervo ótico e disritimia cardíaca. Essa doença causa cegueira bilateral aguda e é principalmente desenvolvida por meninos após os 20 anos de idade (Wallace et al., 1988). Uma mutação por substituição no DNA mitocondrial foi identificado como um fator correlacionado com essa doença em várias famílias. Uma mutação não-sinônima efetiva causada pela substituição de

um **G** (Guanina) para um **A** (Adenina) na segunda posição do códon 340 (C**G**C → C**A**C) do gene **NADH4** (dehidrogenase subunidade 4) muda um aminoácido altamente conservativo, *Arginina*, para uma *Histidina*. Essa mutação ocorre na posição **11.778** do genoma mitocondrial humano. A prevalência [1] da doença de Leber não está claramente definida, no entanto, ela é aceita ser aproximadamente 1:50000 (Orssaud, 2003).

Como estamos estudando as freqüências dos códons nas seqüências de DNA, gostaríamos de responder algumas perguntas, como por exemplo, será que uma mutação na seqüência do gene modifica as freqüências dos códons nesse gene? Essas mutações podem causar alguma doença? Qual a freqüência de mutações sinônimas e não-sinônimas? Qual é o padrão de dependência dos nucleotídeos dentro dos códons? O nucleotídeo que aparece na terceira posição do códon depende do segundo nucleotídeo e do primeiro ou só do imediatamente anterior a ele? Será que a freqüência de um nucleotídeo numa posição depende do nucleotídeo presente numa posição - vizinha? Além disso, também estamos interessados em estudar a variabilidade das seqüências, ou seja, comparar padrões de mutação em seqüências de indivíduos diferentes. Uma maneira exploratória de analisar esses dados fazendo comparações entre as seqüências é comparar várias seqüências do mesmo gene mitocondrial humano com a seqüência de referência de Cambridge desse gene mitocondrial humano.

4.1.1 Comparações entre a SRC e Demais Seqüências

Foram selecionadas 52 seqüências independentes de DNA mitocondrial humano do gene NADH4. Essas seqüências nunca foram analisadas conjuntamente na literatura. Os dados foram obtidos no site do **NCBI** (http://www.ncbi.nlm.nih.gov). O NCBI, ou Centro Nacional para Informação Biotecnológica dos EUA, é considerado o banco de dados central sobre informações genômicas. Vários outros bancos de dados similares estão distribuídos por países da Europa e Japão, mas todos trocam dados em um intervalo de 24 horas com o NCBI. O GenBank é o principal banco de

[1] A prevalência de uma doença na população é a proporção de indivíduos da população que está infectada pela doença.

dados do NCBI e armazena todas as seqüências de DNA (de seqüências pequenas a genomas inteiros), RNA e proteínas disponíveis publicamente. Além do GenBank, que coleta todas as entradas de seqüências, outros bancos do NCBI apresentam as informações organizadas de diferentes maneiras.

Essas seqüências foram comparadas com a SRC e o número de diferenças entre os nucleotídeos da SRC e as demais seqüências foi anotado, ou seja, as seqüências foram comparadas nucleotídeo à nucleotídeo nas 1.377 posições do gene estudado. Esses valores podem ser vistos na Tabela A.15 e A.16 no apêndice. É fácil ver que, na maioria dos casos, as diferenças entre as seqüências ocorrem nas mesmas posições da seqüência, indicando um padrão nas diferenças que pode significar que um determinado aminoácido é codificado por códons diferentes em seqüências diferentes. Porém, esse tipo de diferença entre as seqüências não muda o aminoácido codificado, não alterando assim a cadeia polipeptídica. Mas esse fato é de interesse, pois apesar de não interferir na produção das proteínas, ele altera a freqüência dos códons. Assim, gostaríamos de entender porque em um indivíduo o aminoácido é codificado por um códon e em outro indivíduo, o mesmo aminoácido, na mesma posição da seqüência, no mesmo gene, é codificado por outro códon.

Foram encontradas 176 diferenças em 52 seqüências, ou seja, a média das diferenças encontradas é de 3,38 diferenças por seqüência do gene NADH4 e a mediana é de três nucleotídeos diferentes da SRC por seqüência. Esses valores indicam que é possível traçar um padrão para as seqüências de DNA, como foi feito na seqüência de referência de cambridge, no entanto, é claro que existem diferenças entre a seqüência padrão (SRC) e seqüências de outros indivíduos, por mais que sejam poucas diferenças, são elas que nos fornecem a variabilidade dos dados. A Tabela 4.6 mostra as 263 mutações (diferenças se comparadas com a SRC) encontradas em n = 88 seqüências. Nessa tabela o valor da posição na seqüência se refere à posição dentro do gene NADH4 e não na seqüência inteira do genoma mitocondrial humano.

Analisando a Tabela 4.6 vê-se que foram encontradas 26 mutações diferentes comparando a SRC com as demais. Dentre essas mutações é possível notar que al-

Tabela 4.6: Descrição das Mutações Encontradas comparando-se a SRC com as demais seqüências

Posição	Freq	Códon	Posição no códon	Tipo de Subs.	Subs.	Aminoácido	Mudança
42	9	14	3	transição	ga	Leu	não
105	1	35	3	transição	ct	Ser	não
114	48	38	3	transição	tc	Pro	não
256	2	86	1	transição	tc	Try-Arg	sim
261	1	87	3	transição	ag	Glu	não
325	2	109	1	transição	ag	Tyr-Ala	sim
408	2	136	3	transição	ag	Try	não
456	5	152	3	transição	ct	Tyr	não
540	1	180	3	transição	tc	Thr	não
576	69	192	3	transição	tc	Asp	não
706	2	236	1	transição	tc	Leu	não
723	1	241	3	transição	tc	Tyr	não
771	1	257	3	transição	ga	Met	não
777	12	259	3	transição	ct	Tyr	não
888	11	296	3	transição	ct	Leu	não
900	1	300	3	transição	ct	Ala	não
942	2	314	3	transição	tc	Iso	não
960	73	320	3	transição	ga	Gly	não
1.003	1	335	1	transição	ga	Glu-Lys	sim
1.019	9	340	2	transição	ga	Arg-His	sim
1.155	5	385	3	transição	ga	Thr	não
1.185	1	395	3	transição	tc	Leu	não
1.248	1	416	3	transição	ga	Tyr	não
1.267	1	423	1	transição	ag	Iso-Val	sim
1.325	1	442	2	transição	ct	Ser-Phe	sim
1.333	1	445	1	transversão	ca	Leu-Iso	sim

gumas acontecem com uma freqüência muito maior do que outras. Esse é o caso das mutações encontradas nas posições 960, 576 e 114 do gene NADH4 do genoma mitocondrial humano. Algumas mutações apresentam uma característica interessante, são não sinônimas, ou seja, essa mutação provoca uma mudança no aminoácido codificado, que provoca mudança na síntese de proteínas, podendo assim causar alguma doença. Como já era esperado, as mutações na primeira e segunda posições

do códon são pouco frequentes e as da terceira posição são muito frequentes. Além disso, é possível ver que as transições ocorrem muito mais frequentemente do que transversões. Observe que temos nove indivíduos com a doença de Leber nessa amostra, ou seja, as nove mutações encontradas na posição 1.019 que provocam a mudança do aminoácido Arginina pra uma Histidina causam a doença de Leber.

4.2 Ajuste de Modelos para a Seqüência de Referência de Cambridge

Inicialmente usamos apenas uma seqüência de DNA mitocondrial para analisar as freqüências dos códons. As freqüências observadas dos códons da SRC para o gene NADH4 encontram-se no apêndice na Tabela A.17.

Os códons de parada para os genes mitocondriais humanos são: **TAA**, **TGA**, **AGG** e **AGA**. Três covariáveis foram utilizadas no estudo: **AARISK**, uma medida do risco de mutação de um aminoácido, **AVDIST**, que mede quão típico é um aminoácido e **TSCORE**, que mede a proximidade dos códons com os códons de parada. AARISK é uma variável que mede o risco de ocorrer mutação entre dois códons (que não sejam códons de parada) que codificam aminoácidos com propriedades físicas e químicas bem diferentes. Quando ocorre mutação entre códons que codificam o mesmo aminoácido, tem-se a mutação sinônima.

Existem diferentes métodos para obter as distâncias entre os aminoácidos, no entanto, a usada para calcular as variáveis AARISK e AVDIST está em Granthan(1974). A fórmula para calcular as distâncias entre os aminoácidos é baseada em três propriedades químicas dos aminoácidos: a composição (c), a polaridade (p) e o volume molecular (v). Assim, as distâncias são dadas por:

$$D_{ij} = [\alpha(c_i - c_j)^2 + \beta(p_i - p_j)^2 + \gamma(v_i - v_j)^2]^{1/2},$$

em que α, β e γ são os quadrados do inverso das médias, respectivamente da composição, polaridade e volume molecular dos 20 aminoácidos (Granthan, 1974).

Assim, AARISK é a distância média ponderada entre um dado aminoácido e os outros 19 aminoácidos. Em geral, AARISK é diferente para códons sinônimos. AVDIST é a distância média de Granthan entre um dado aminoácido e os outros sem considerar a ponderação. Quanto menor o valor de AVDIST, mais típico é o aminoácido codificado. Essa variável é constante entre códons sinônimos. A variável TSCORE é derivada dos códons, ou seja, é o número de mudanças únicas nas bases que causaria a mutação entre um códon e um códon de parada, e ela pode assumir os valores, 0,1 ou 2. Por exemplo, o códon TTA tem um TSCORE igual a um, pois mudando o segundo T para um A, obtém-se o códon de parada TAA. O TSCORE de TGA é 2, pois mudando T para A ou G para A, obtém-se os códons de parada AGA ou TAA, respectivamente. O códon CCC (ou qualquer outro códon com dois C's) tem um TSCORE igual a zero, pois se ocorre mutação em um C, restam ainda um ou dois C's e nenhum dos códons de parada tem C como uma das bases que os formam, então é impossível para uma substituição única de base gerar um códon de parada nesse caso. Como no caso de AARISK, TSCORE também pode ser diferente entre códons sinônimos. As covariáveis não possuem valores para os códons de parada, por exemplo, para a covariável TSCORE, não tenho como contar o número de mudanças necessárias para um códon de parada se transformar em um códon de parada, pois ele já é um códon de parada. Os valores das covariáveis para os 60 códons efetivos podem ser vistos na Tabela A.17 no apêndice. Assim, é bom ressaltar que em todos os modelos que incluem covariáveis são utilizadas 60 freqüências e não 64.

4.2.1 Teste da Razão de Verossimilhanças para o Modelo Aditivo

O número de parâmetros do modelo aditivo (3.4) é muito grande e nesse modelo consideramos parâmetros diferentes para a mesma variável nos logitos da segunda e terceira posições, por esse motivo, vamos implementar um modelo com um número de parâmetros menor, ou seja, considerando o mesmo parâmetro para a mesma variável em logitos diferentes. Depois utilizaremos o teste da Razão de Verossimilhanças (RV) para averiguar se os parâmetros a mais no modelo maior são realmente

necessários. A estatística do teste da RV é dada por:

$$RV \;=\; -2\,log\,\left(\frac{L_1}{L_0}\right) \sim \chi^2_{(p_1-p_0)}$$

em que p_1 é o número de parâmetros do modelo L_1 e p_0 é o número de parâmetros do modelo L_0.

Para entender melhor e poder visualizar as hipóteses do teste, os parâmetros envolvidos no teste foram colocados em negrito nos logitos, explicando assim a diferença entre os modelos. Os logitos do modelo aditivo com 21 parâmetros são dados por:

$$\begin{aligned}
\theta_{11} &= \alpha_{11} + \beta_1\, X_1 + ... + \beta_p\, X_p \\
\theta_{12} &= \alpha_{12} + \beta_1\, X_1 + ... + \beta_p\, X_p \\
\theta_{13} &= \alpha_{13} + \beta_1\, X_1 + ... + \beta_p\, X_p \\
\theta_{21} &= \alpha_{21} + \gamma_{11}Z_{11} + \gamma_{12}Z_{12} + \gamma_{13}Z_{13} + \beta_1\, X_1 + ... + \beta_p\, X_p \\
\theta_{22} &= \alpha_{22} + \gamma_{11}Z_{11} + \gamma_{12}Z_{12} + \gamma_{13}Z_{13} + \beta_1\, X_1 + ... + \beta_p\, X_p \\
\theta_{23} &= \alpha_{23} + \gamma_{11}Z_{11} + \gamma_{12}Z_{12} + \gamma_{13}Z_{13} + \beta_1\, X_1 + ... + \beta_p\, X_p \\
\theta_{31} &= \alpha_{31} + \boldsymbol{\gamma_1}Z_{11} + \boldsymbol{\gamma_2}Z_{12} + \boldsymbol{\gamma_3}Z_{13} + \gamma_{21}Z_{21} + \gamma_{22}Z_{22} + \gamma_{23}Z_{23} + \beta_1\, X_1 + ... + \beta_p\, X_p \\
\theta_{32} &= \alpha_{32} + \boldsymbol{\gamma_1}Z_{11} + \boldsymbol{\gamma_2}Z_{12} + \boldsymbol{\gamma_3}Z_{13} + \gamma_{21}Z_{21} + \gamma_{22}Z_{22} + \gamma_{23}Z_{23} + \beta_1\, X_1 + ... + \beta_p\, X_p \\
\theta_{33} &= \alpha_{33} + \boldsymbol{\gamma_1}Z_{11} + \boldsymbol{\gamma_2}Z_{12} + \boldsymbol{\gamma_3}Z_{13} + \gamma_{21}Z_{21} + \gamma_{22}Z_{22} + \gamma_{23}Z_{23} + \beta_1\, X_1 + ... + \beta_p\, X_p
\end{aligned}$$

Os logitos para o modelo aditivo menor são dados por:

$$
\begin{aligned}
\theta_{11} &= \alpha_{11} + \beta_1 X_1 + ... + \beta_p X_p \\
\theta_{12} &= \alpha_{12} + \beta_1 X_1 + ... + \beta_p X_p \\
\theta_{13} &= \alpha_{13} + \beta_1 X_1 + ... + \beta_p X_p \\
\theta_{21} &= \alpha_{21} + \gamma_{11}Z_{11} + \gamma_{12}Z_{12} + \gamma_{13}Z_{13} + \beta_1 X_1 + ... + \beta_p X_p \\
\theta_{22} &= \alpha_{22} + \gamma_{11}Z_{11} + \gamma_{12}Z_{12} + \gamma_{13}Z_{13} + \beta_1 X_1 + ... + \beta_p X_p \\
\theta_{23} &= \alpha_{23} + \gamma_{11}Z_{11} + \gamma_{12}Z_{12} + \gamma_{13}Z_{13} + \beta_1 X_1 + ... + \beta_p X_p \\
\theta_{31} &= \alpha_{31} + \gamma_{11}Z_{11} + \gamma_{12}Z_{12} + \gamma_{13}Z_{13} + \gamma_{21}Z_{21} + \gamma_{22}Z_{22} + \gamma_{23}Z_{23} + \beta_1 X_1 + ... + \beta_p X_p \\
\theta_{32} &= \alpha_{32} + \gamma_{11}Z_{11} + \gamma_{12}Z_{12} + \gamma_{13}Z_{13} + \gamma_{21}Z_{21} + \gamma_{22}Z_{22} + \gamma_{23}Z_{23} + \beta_1 X_1 + ... + \beta_p X_p \\
\theta_{33} &= \alpha_{33} + \gamma_{11}Z_{11} + \gamma_{12}Z_{12} + \gamma_{13}Z_{13} + \gamma_{21}Z_{21} + \gamma_{22}Z_{22} + \gamma_{23}Z_{23} + \beta_1 X_1 + ... + \beta_p X_p
\end{aligned}
$$

Estamos interessados em testar as seguintes hipóteses:

$$
\begin{aligned}
\boldsymbol{H_0} \; : \; & \gamma_{11} = \gamma_1 \quad \text{vs} \quad \boldsymbol{H_1} : \gamma_{11} \neq \gamma_1 \\
& \gamma_{12} = \gamma_2 \qquad\qquad \gamma_{12} \neq \gamma_2 \\
& \gamma_{13} = \gamma_3 \qquad\qquad \gamma_{13} \neq \gamma_3
\end{aligned}
$$

Foram rodados quatro modelos aditivos para o gene NADH4 do genoma mitocondrial humano, utilizando-se a SRC. Os resultados podem ser vistos na Tabela 4.7.

Tabela 4.7: Teste da Razão de Verossimilhanças (RV)

Modelos Aditivos	-log L	RV
L_1. Com Covariáveis (21 parâmetros)	117,56	
L_0. Com Covariáveis (18 parâmetros)	126,56	18
L_1. Sem Covariáveis (18 parâmetros)	118,46	
L_0. Sem Covariáveis (15 parâmetros)	132,46	28

O valor tabelado da $\chi^2_{(3;0,05)}$ é 7,815, assim, rejeita-se a hipótese nula do teste,

ou seja, para os modelos com covariáveis, o modelo aditivo completo (com mais parâmetros) é mais adequado, ou seja, os modelos com parâmetros diferentes para os coeficientes dos Z's nos logitos da segunda e da terceira posições do códon são mais adequados. Assim, de agora em diante usaremos apenas o modelo aditivo com o maior número de parâmetros.

4.2.2 Comparação entre os Modelos

Os outros modelos com e sem estrutura de dependência também foram rodados e os resultados podem ser vistos na Tabela 4.8.

Tabela 4.8: Ajuste dos Modelos para o Gene NADH4, SRC

Modelos	n° parâ	-2 log L	AIC
1. Aditivo Com Covariáveis	21	235,12	277,12
2. Aditivo Sem Covariáveis	18	236,92	272,92
3. Markov Com Covariáveis	18	235,32	**271,32**
4. Markov Sem Covariáveis	15	243,32	273,32
5. Igual. Pred. Com Cov	18	248,52	284,52
6. Igual. Pred. Sem Cov	15	250,12	280,12
7. Independente Com Covariáveis	12	298,52	322,52
8. Independente Sem Covariávies	9	302,32	320,32

Observando a Tabela 4.8 vemos que segundo o critério de Akaike (AIC) o melhor modelo é o de Markov com covariáveis (3). No entanto, para testar se esse modelo é realmente o mais parcimonioso, gostaríamos de utilizar o teste da RV para comparar os modelos. Comparando o modelo aditivo com covariáveis (1) com o modelo aditivo sem covariáveis (2), temos RV = 1,8 com 3 g.l., ou seja, como o valor tabelado da χ^2 com 3 g.l. é 7,815, não rejeitamos a hipótese de que os coeficientes das covariáveis são nulos. Assim, concluímos que, nesse caso, as covariáveis não são importantes no modelo aditivo. Comparando o modelo aditivo com covariáveis (1) com o modelo de Markov com covariáveis (3), temos RV = 0,2 com 3 g.l., ou seja, a estrutura de dependência utilizada no modelo de Markov é mais adequada do que a utilizada no modelo aditivo. Assim, nota-se que uma estrutura Markoviana

de primeira ordem é suficiente para explicar a dependência dos nucleotídeos nos códons. Comparando o modelo aditivo com covariáveis (1) com o modelo igualmente preditivo com covaríáveis (5), temos RV = 13,4 com 3 g.l., ou seja, a estrutura de dependência utilizada no modelo aditivo é mais adequada do que a estrutura do modelo igualmente preditivo. Para a comparação entre o modelo aditivo com covariáveis (1) e o modelo independente com covariáveis (7), temos RV = 63,4 com 9 g.l., ou seja, como o valor tabelado da χ^2 com 9 g.l. e nível de significância 0,05 é igual a 16,919, concluímos que é mais adequado considerar uma estrutura de dependência para os nucleotídeos do que não considerar nenhuma estrutura de dependência. O interessante agora é comparar os modelos de Markov com e sem covariáveis para averiguar qual é o mais parcimonioso. Assim, o valor da estatística do teste, RV = 8 com 3 g.l., ou seja, rejeita-se a hipótese nula de que os coeficientes das covariáveis sejam nulos ao nível de 5%. Assim, está confirmada a importância das covariáveis nos modelos com estruturas Markovianas de dependência .

Comparando-se os modelos igualmente preditivos com e sem covariáveis, obtemos RV = 1,68 com 3 g.l., o que significa que não se rejeita a hipótese de que os coeficientes das covariáveis são nulos. Analogamente, nos modelos independentes, o valor da estatísica RV = 3,8 com 3 g.l., que também não rejeita a hipótese nula. Então, conclui-se que, em geral, os modelos que não continham covariáveis são mais parcimoniosos do que os modelos com as covariáveis. É possivel perceber também que o pior modelo segundo o critério do AIC, é o modelo independente com covariáveis, confirmando nossa suposição de que uma estrutura de dependência é necessária para explicar as freqüências dos códons. Vamos agora apresentar as estimativas dos parâmetros para o modelo mais parcimonioso encontrado para a SRC, ou seja, o modelo de Markov com covariáveis. Na Tabela 4.9, β_1, β_2 e β_3 representam respectivamente, os coeficientes das variáveis AARISK, TSCORE e AVDIST.

Para fazer a interpretação dos parâmetros é bom lembrar que em (3.15) θ_{ij}, $i =$ 1,2 e 3, representa o logito da i-ésima posição do códon das as bases C, A e G contra T, respectivamente, por exemplo, θ_{11} representa o logito de C contra T na primeira posição do códon.

Tabela 4.9: Estimativa dos Parâmetros do Modelo de Markov com Covariáveis (3)

Parâmetros	Estimativas	exp(estimativas)
α_{11}	3,0935	22,05
α_{12}	3,4032	30,06
α_{13}	2,5050	12,24
α_{21}	2,4850	12,00
α_{22}	2,4621	11,73
α_{23}	2,3876	10,89
α_{31}	3,2634	26,14
α_{32}	3,2763	26,48
α_{33}	0,9344	2,55
γ_{21}	-1,4829	0,23
γ_{22}	-0,2350	0,79
γ_{23}	0,4818	1,62
γ_{31}	0,1556	1,17
γ_{32}	2,3649	10,64
γ_{33}	2,8562	17,40
β_1	-0,0223	0,98
β_2	-0,4191	0,66
β_3	-0,1571	0,85

Analisando a Tabela 4.9 podemos ver que o fato da base C aparecer na primeira posição do códon diminui a chance de aparecerem os nucleotídeos A, C ou G na segunda posição do códon em 4,39 (1/0,23) vezes, ou seja, a chance da base T aparecer na segunda posição do códon é 4,39 vezes maior do que aparecerem as bases A, C ou G quando sabemos que o nucleotídeo C apareceu na primeira posição. A chance da base T aparecer na segunda posição do códon quando a base A apareceu na primeira posição do códon é 1,26 (1/0,79) vezes maior do que aparecerem as bases A, C ou G. A chance de aparecerem os nucleotídeos A, C ou G na segunda posição do códon é 1,62 vezes maior do que aparecer um T quando um G apareceu na primeira posição. A chance dos nucleotídeos A, C ou G aparecerem na terceira posição do códon quando um C apareceu na segunda posição do códon é 1,17 vezes maior do que aparecer um T na terceira posição. A chance de qualquer um dos nucleotídeos A, C ou G aparecerem na terceira posição do códon quando um A apareceu na segunda

posição é 10,64 vezes maior do que aparecer um T na terceira posição. A chance de qualquer um dos nucleotídeos A, C ou G aparecerem na terceira posição do códon quando um G apareceu na segunda posição é 17,4 vezes maior do que aparecer um T.

Para fazer a interpretação dos interceptos é bom lembrar que eles representam um incremento no valor dos logitos. Assim, por exemplo, para os logitos da terceira posição vemos que independentemente do nucleotídeo que aparece na segunda posição, a chance de aparecer um A ou um C na terceira posição é bem maior do que aparecer um G, pois os logitos θ_{31} e θ_{32} são bem maiores do que θ_{33}. Além disso, podemos dizer que o nucleotídeo A tem chance maior de ser encontrado na primeira posição do que nas outras posições do códon, já que o logito θ_{12} é maior do que os logitos θ_{22} e θ_{32}. Analogamente, podemos dizer que o nucleotídeo C tem mais chance de ser encontrado na terceira posição do códon do que nas outras posições.

Analisando os coeficientes das covariáveis percebemos que β_2 é negativo, ou seja, o coeficiente da variável TSCORE é negativo, o que indica que quanto maior for o número de mudanças de base única necessárias para o códon mutar e se transformar em um códon de parada, menor será a frequência das bases A, C e G se comparadas com a base T.

Analisando a Tabela 4.10, vemos que na primeira posição do códon o nucleotídeo que aparece com maior frequência é o A. Analisando as frequências relativas para a segunda posição do códon, vemos que o nucleotídeo T é o que aparece mais. Para a terceira posição observamos que o nucleotídeo que mais aparece é o C e o nucleotídeo que praticamente não aparece nessa posição é o G.

4.3 Ajuste dos Modelos para várias Seqüências de DNA

Até agora, usamos apenas uma seqüência de DNA, a SRC, que é aceita e utilizada mundialmente como referência padronizada de seqüência humana.

Tabela 4.10: Frequências Relativas das bases por posição no códon para a SRC

Probabilidades	Freq. Relativa
p_{11}	0,309
p_{12}	0,344
p_{13}	0,148
p_{14}	0,198
p_{21}	0,283
p_{22}	0,179
p_{23}	0,113
p_{24}	0,428
p_{31}	0,435
p_{32}	0,383
p_{33}	0,037
p_{34}	0,144

Seria interessante utilizar mais do que uma única seqüência de DNA como amostra para que se tenha resultados mais realistas, pois assim estaríamos realmente avaliando a variabilidade dos dados. Assim, os oito modelos foram rodados novamente, só que agora usando várias seqüências de DNA, ou seja, utilizamos a mesma seqüência utilizada anteriormente (gene NADH4) na SRC, mas agora, usamos seqüências de vários indivíduos diferentes. As seqüências podem ser consideradas independentes, pois provêem de indivíduos epidemiologicamente inpedententes, ou seja, os indivíduos foram selecionados ao acaso (não existem membros de uma mesma família na amostra).

Foram utilizadas 30 seqüências de DNA, que também foram retiradas so site do NCBI. Desses 30 indivíduos, sete tem a doença de Leber, oito são normais, quatro tem o Mal de Alzheimer, um tem Diabetes, oito tem o Mal de Parkinson, um é obeso e um é a seqüência da referência de cambridge.

Analisando a Tabela 4.11 é possível ver que sob o critério de Akaike o melhor modelo é o aditivo com covariáveis (1) pois é o que possui o menor valor do AIC. Comparando o modelo aditivo com covariáveis (1) com o modelo aditivo sem covariáveis (2) através do teste da RV, obtemos o valor da estatística do teste, RV =

Tabela 4.11: Ajuste dos Modelos para o Gene NADH4, n=30 seqüências

Modelos	n° parâ	-2 log L	AIC
1. Aditivo + Covariáveis	21	5.627,36	**5.669,36**
2. Aditivo - Covariáveis	18	7.041,56	7.077,56
3. Markov + Covariáveis	18	6.172,16	6.208,16
4. Markov - Covariáveis	15	7.243,76	7.273,76
5. Igual. Pred. + Cov	18	6.211,76	6.247,76
6. Igual. Pred. - Cov	15	7.465,76	7.495,76
7. Independente + Cov	12	8.203,76	8.227,76
8. Independente - Cov	9	9.057,56	9.075,56

1.414,2 com 3 graus de liberdade, ou seja, rejeitamos a hipótese nula de que os coeficientes das covariáveis são nulos e assim corfirmamos a importância das covariáveis nesse modelo. Comparando o modelo aditivo com covariáveis (1) com o modelo de Markov com covariáveis (3) obtemos RV = 544,8 com 3 g.l., o que significa que o modelo aditivo é mais adequado para estabelecer uma estrutura de dependência para os dados. Comparando o modelo aditivo com covariáveis (1) com o modelo igualmente preditivo (5), obtemos RV = 584,4 com 3 g.l., ou seja, o padrão de dependência utilizado no modelo aditivo é mais adequado do que o utilizado no modelo igualmente preditivo. Por último, comparando o modelo aditivo com covariáveis (1) com o modelo independente com covariáveis (7), obtemos RV = 2.576,4 com 9 g.l., ou seja, a hipótese de que uma estrutura de dependência não é necessária é rejeitada, assim confirmamos a suposição de que uma estrutura de dependência nos dados é importante. Para testar a importância das covariáveis nos modelos, não apenas no aditivo, mas em todos os outros também, foram feitos testes da RV de acordo com as seguintes hipóteses:

$$H_0 : \beta_1 = 0, \ ..., \ \beta_p = 0 \quad \text{vs} \quad H_1 : \textit{pelo menos um } \beta \neq 0.$$

Para o modelo de Markov o valor da estatística, RV é 1.071,6 com 3 graus de liberdade, o que significa que rejeita-se a hipótese nula, ou seja, as covariáveis são significativas para o modelo de Markov. No modelo igualmente preditivo, obtemos RV = 1.254 também com três graus de liberdade, o que significa que o modelo com as

covariáveis é mais parcimonioso do que o modelo sem as covariáveis. Analogamente, para o modelo independente as covariáveis também se mostraram importantes com um valor RV = 853,8. Além disso, observando os valores do critério de Akaike, vemos que todos os modelos com covariáveis possuem AIC menores do que seus similares sem covariáveis, o que indica que os modelos com covariáveis são mais parcimoniosos. Assim, podemos concluir que quando o tamanho amostral aumenta o modelo selecionado como o modelo mais parcimonioso muda, ou seja, é necessário supor que o nucleotídeo que aparece na terceira posição do códon depende dos nucleotídeos que aparecem na primeira e segunda posições e não somente da posição imediatamente anterior, que é o que acontece no modelo de Markov. Sob o critério de Akaike, o pior modelo é o independente sem covariáveis, pois possui o maior AIC. Esse fato já era esperado, pois esse modelo assume independência entre as posições do códon e essa suposição não é muito realista do ponto de vista biológico.

Pode-se concluir que, para analisar dados de freqüências de códons mitocondriais, é aconselhável usar uma estrutura de dependência entre os nucleotídeos, como por exemplo, um modelo aditivo. Os modelos que assumem independência entre os nucleotídeos podem ser equivocados, já que não são os mais parcimoniosos.

Como já era esperado, a importância das covariáveis no modelo foi confirmada, ou seja, independente do modelo escolhido para analisar as freqüências dos códons, a inclusão das covariáveis mostrou uma melhora significativa no modelo.

É interessante notar que a Tabela 4.11 apresenta resultados mais plausíveis com as suposições estabelecidas nos capítulos anteriores do que a Tabela 4.8, pois essa nova tabela foi gerada utilizando várias seqüências, conseguindo assim extrair a variabilidade dos dados. Achamos que uma única seqüência pode fornecer o caminho a ser seguido, um indício da resposta esperada, no entanto, uma amostra aleatória de seqüências é o caminho correto a seguir quando desejamos fazer inferências.

Desejamos agora rodar o modelo aditivo novamente, utilizando várias combinações das covariáveis, para averiguar qual combinação forma o modelo mais parcimonioso. Ainda utilizando 30 seqüências, os modelos foram rodados e os resultados

podem ser vistos na Tabela 4.12.

Tabela 4.12: Ajuste do Modelo Aditivo para várias combinações de covariáveis, n=30 seqüências

Variáveis	n° parâ	-2 log L	AIC
1. AARISK	19	7.965,76	8.003,76
2. TSCORE	19	8.061,76	8.099,76
3. AVDIST	19	6.385,76	6.423,76
4. AARISK-AVDIST	20	7.795,76	7.835,76
5. AARISK-TSCORE	20	5.725,76	5.765,76
6. TSCORE-AVDIST	20	6.231,76	6.271,76
7. AARISK-TSCORE-AVDIST	21	5.627,36	**5.669,36**
8. Nenhuma	18	7.041,56	7.077,56

Analisando a Tabela 4.12, vemos que o modelo aditivo com todas as covariáveis é o mais parcimonioso, pois possui o menor AIC. Assim, as estimativas dos parâmetros do modelo mais parcimonioso, segundo o critério do AIC, estão apresentadas na Tabela 4.13.

Considerando homogeneidade ao longo das seqüências e observando a Tabela 4.13 vemos que a chance de aparecer um T na segunda posição do códon dado que apareceu um C na primeira posição é 5(1/0,2) vezes maior do que aparecerem os nucleotídeos A ou C ou G. A chance de aparecer um T na segunda posição do códon é 3,36 vezes maior do que aparecerem os nucleotídeos A ou C ou G quando o nucleotídeo A apareceu na primeira posição. A chance de aparecerem os nucleotídeos A ou C ou G na terceira posição do códon é 3 vezes maior do que parecer um T quando o nucleotídeo A apareceu na primeira posição do códon. Outra interpretação importante é a do γ_3, ou seja, a chance de aparecerem os nucleotídeos A ou C ou G na terceira posição do códon é 6,39 vezes maior do que aparecer o nucleotídeo T quando o nucleotídeo G apareceu na primeira posição. Analogamente ao que acontece no modelo de Markov para a SRC, a chance de aparecer um G em qualquer uma das três posições no códon é menor do que aparecer os nucleotídeos A ou C, esse fato deve-se aos valores dos interceptos α_{13}, α_{23} e α_{33} serem pequenos se comparados com os outros interceptos para a mesma posição.

Tabela 4.13: Estimativa dos parâmetros do Modelo Aditivo com Covariáveis (1)

Parâmetros	Estimativas	exp(estimativas)
α_{11}	7,5429	1.887,29
α_{12}	7,6494	2.099,39
α_{13}	6,8056	902,89
α_{21}	7,4074	1.648,14
α_{22}	6,9491	1.042,21
α_{23}	6,4873	656,75
α_{31}	8,1015	3.299,41
α_{32}	7,9735	2.903,00
α_{33}	5,5722	263,01
γ_{11}	-1,6045	0,20
γ_{12}	-1,2126	0,30
γ_{13}	0,7817	2,19
γ_{21}	0,6360	1,89
γ_{22}	-0,8559	0,42
γ_{23}	0,4903	1,63
γ_1	-0,2677	0,77
γ_2	1,1168	3,06
γ_3	1,8545	6,39
β_1	0,0315	1,03
β_2	0,2082	1,09
β_3	-0,9629	0,38

4.3.1 Diagnóstico dos Modelos

Em regressão linear, as medidas de distâncias entre valores observados e ajustados, bem como diagnósticos para avaliar os efeitos de observações no ajuste dos dados, são funções dos resíduos. Em regressão logística binária existem várias medidas possíveis para calcular as diferenças entre os valores observados e ajustados, como por exemplo, os resíduos de Pearson, que fornecem a estatística χ^2 de Pearson e o resíduo deviance (Hosmer & Lemeshow, 1989). No entanto, para regressão logística politômica, as categorias múltiplas de resposta complicam o problema se comparado com o modelo de regressão logística binária onde temos apenas um valor ajustado. Lesaffre (1986) propôs extensões de testes de adequação de ajuste e diagnósticos para

modelos de regressão logística politômica. No entanto, esses métodos não são facilmente calculados e são computacionalmente inviáveis. Hosmer & Lemeshow (1989) recomendam calcular as medidas de diagnóstico utilizadas para regressão logística binária usando o ajuste da regressão logística individual para cada logito e depois integrar os resultados.

Gostaríamos de fazer uma análise de diagnóstico dos modelos ajustados, para averiguar qual é o modelo que melhor se ajusta aos dados. Assim sendo, gostaríamos de calcular os resíduos dos modelos para averiguar a heterocedasticidade dos dados e a possível presença de valores aberrantes. No entanto, os resíduos para o modelo logístico politômico não são muito realistas quando aplicados a dados de seqüências de DNA. Resolvemos então utilizar uma abordagem alternativa para calcular uma medida que pudesse se aproximar dos resíduos do modelo, ou seja, comparando as probabilidades observadas dos códons com as probabilidades estimadas. Assim, o que estamos chamando de resíduo de agora em diante não é o resíduo padrão conhecido em regressão linear simples (Y_i - $\hat{Y}_i$). O que estamos chamando de "resíduos", são as diferenças entre as probabilidades observadas e ajustadas dos códons, ou seja, P_i - $\hat{P}_i$, em que i reprenta o códon (i = 1,2, ..., 64). Com o intuito de comparar os oito modelos ajustados aos dados, calculamos a soma dos quadrados dos resíduos (SQR). É bom lembrar que aqui estamos comparando as probabilidades dos códons e não poderemos fazer análise de resíduos como no modelo linear geral, em que os resíduos tem distribuição normal. Os resultados podem ser vistos na tabela 4.14.

Tabela 4.14: SQR para os modelos ajustados para n= 30 seqüências

Modelos	SQR
1. Aditivo com Covariáveis	0,0051
2. Aditivo sem Covariáveis	**0,0028**
3. Markov com Covariáveis	0,0056
4. Markov sem Covariáveis	0,0032
5. Igual. Pred. com Covariáveis	0,0059
6. Igual. Pred. sem Covariáveis	0,0031
7. Independente com Covariáveis	0,0088
8. Independente sem Covariáveis	0,0061

Analisando a Tabela 4.14 observamos que o melhor ajuste dos dados é dado pelo modelo aditivo sem covariáveis, pois é o que possui menor SQR. Essa análise dos resíduos é apenas descritiva, para se ter uma idéia do comportamento dos modelos. Os gráficos dos resíduos para os oito modelos ajustados encontram-se nas figuras 4.7, 4.8, 4.9 e 4.10.

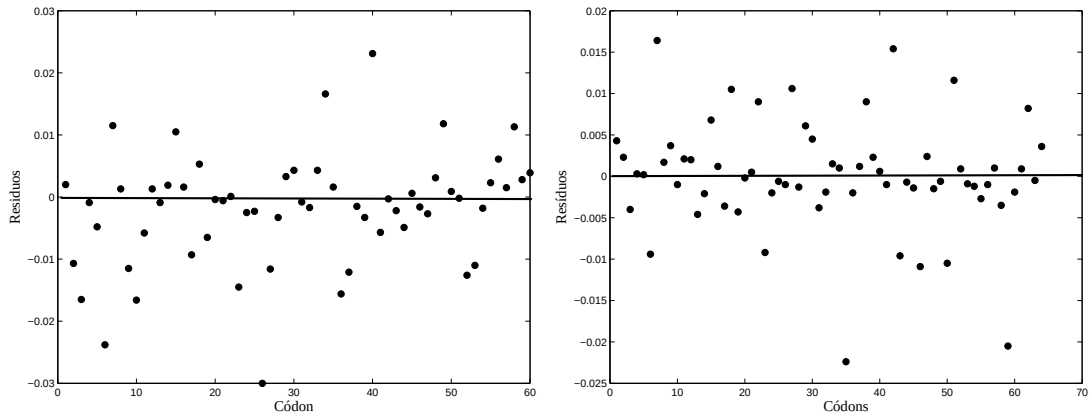

Figura 4.7: Modelo Aditivo com e sem Covariáveis, n = 30 seqs.

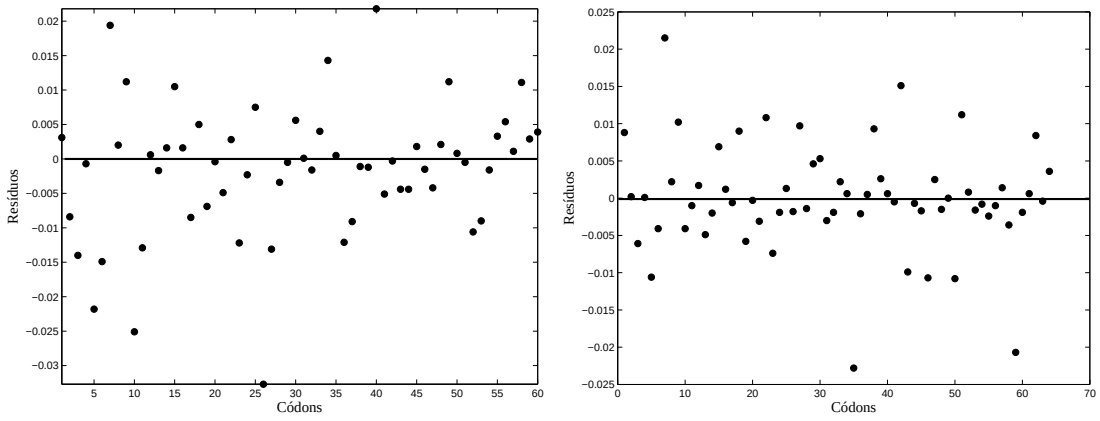

Figura 4.8: Modelo de Markov com e sem Covariáveis, n = 30 seqs.

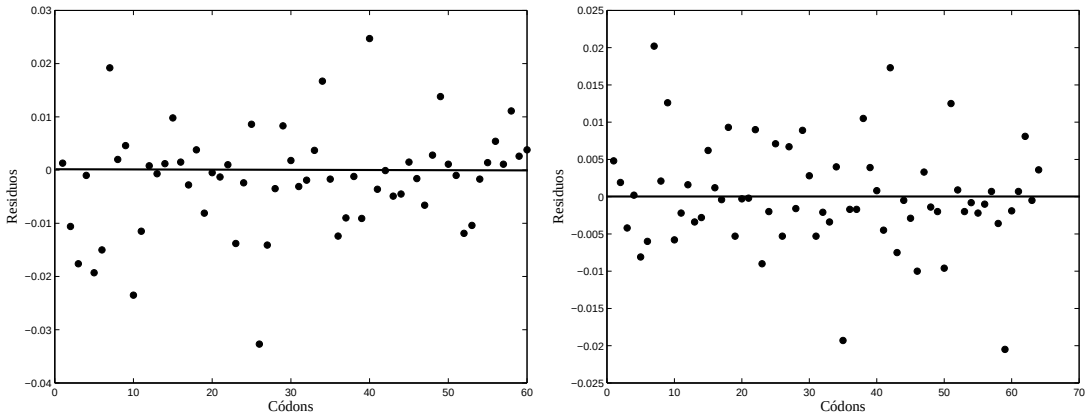

Figura 4.9: Igualmente Preditvo com e sem Covariáveis, n = 30 seqs.

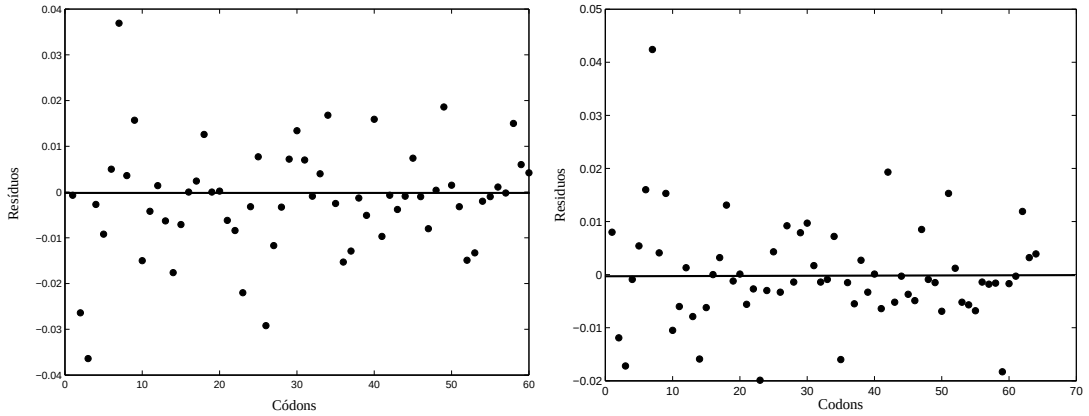

Figura 4.10: Independente com e sem Covariáveis, n = 30 seqs.

Fazendo uma análise puramente visual dos gráficos dos resíduos para os oito modelos, observa-se que em todos os modelos os resíduos estão distribuídos aleatoriamente em torno de zero, não apresentando nenhum padrão de heterogeneidade, nem valores aberrantes.

Podemos concluir que analisando os modelos para saber qual é o mais parcimonioso, utilizando o critério de Akaike e o teste da RV, vemos que o modelo aditivo com covariáveis é escolhido. No entanto, quando fazemos os diagnósticos dos modelos para saber qual é o modelo que melhor se adequa aos dados, vemos que o modelo

aditivo sem covariáveis parece ser o melhor. Assim, podemos dizer que, de um modo geral, a classe dos modelos aditivos são adequados para analisar freqüências de códons em seqüências de DNA.

"A questão de interpretação dos parâmetros é crucial num modelo logístico, implicando que uma forma puramente mecânica de seleção de modelos pode levar a um modelo sem sentido e de difícil interpretação. Muitas vezes, variáveis consideradas biologicamente importantes não devem ser deixadas de lado pela sua falta de significância estatística. Assim, a seleção de um modelo logístico deve ser um proceso conjugado de seleção estatística de modelos e bom senso"(Paula, 2004). Assim, sugerimos o modelo aditivo com covariáveis como uma escolha adequada de modelo para analisar as freqüências dos códons em seqüências de DNA, pois esse modelo, além de ter uma estrutura de dependência adequada entre os nucleotídeos, inclui variáveis biologicamente importantes para o estudo.

As probabilidades observadas e ajustadas para todos os códons para os modelos com e sem covariáveis podem ser vistas, respectivamente, nas Tabelas A.18 e A.20 no apêndice.

5 Considerações Finais

De acordo com as análises feitas no capítulo 4, foi visto que, utilizando apenas uma seqüência de DNA, a estrutura de dependência adequada para os nucleotídeos dentro do códon seria uma estrutura Markoviana de primeira ordem. No entanto, quando várias seqüências de DNA são utilizadas na amostra, a estrutura de dependência adequada é a do modelo aditivo. Esse fato reforça a idéia de que seria interessante utilizar mais do que uma seqüência de DNA nas análises para se ter uma idéia global das freqüências dos códons em vários indivíduos distintos.

Assim, conclui-se que quando se deseja estudar as freqüências dos códons em seqüências de DNA é interessante utilizar estrutura de dependência entre os nucleotídeos no códon. Se essa estrutura é aditiva, o nucleotídeo que aparece na segunda posição do códon depende do nucleotídeo que apareceu na primeira e aquele que aparece na terceira posição depende dos nucleotídeos que apareceram na primeira e segunda posiçôes do códon.

Essa conclusão estatística reforça a suposição de que códons aparecem com freqüências desiguais em seqüências codificadoras, pois existe uma relação de dependência entre os nucleotídeos que formam o códon, assim, um códon específico pode aparecer muito mais vezes na sequüência do que um outro códon sinônimo, pois, mesmo codificando o mesmo aminoácido, a composição dos nucleotídeos que formam os códons pode determinar que ele aparece com uma freqüência maior .

Nesse trabalho, utilizamos seqüências de indivíduos com a doença de Leber para enfatizar que os indivíduos da amostra são independentes e para estudar as

diferenças nas freqüências dos códons quando sabemos que existem indivíduos com mutações na seqüência que causam doenças. Como a prevalência da doença de Leber é muito pequena, por se tratar de uma doença rara, não foram observadas grandes diferenças nas freqüências dos códons de indivíduos doentes e sadios. No entanto, para estudos futuros seria interessante utilizar uma doença causada por mutação genética, que tenha uma grande prevalência (câncer de mama, por exemplo) para que se possa visualizar essas diferenças entre os indivíduos sadios e doentes, bem como desenvolver métodos de estimação que leve em conta a prevalência da doença.

Os modelos propostos por Bonney et al. (1994) são inovadores pois utilizam variáveis explicativas e estrutura de dependência para explicar os dados das freqüências dos códons. No entanto, esses modelos supõem independência entre os códons na seqüência de DNA, ou seja, a estrutura de dependência está simplesmente nos nucleotídeos dentro dos códons. Biologicamente faz mais sentido pensar em uma estrutura de dependência mais global, pois a suposição de independência entre os códons é considerada forte, ou seja, biologicamente irreal.

Como já visto na revisão de conceitos biológicos, a ocorrência de mutação na terceira posição do códon é muito mais frequente do que nas outras posições do códon, pois, em muitos casos, uma alteração de base na terceira posição não muda o aminoácido codificado. O mesmo não ocorre para mudanças de base na primeira e segunda posição do códon, já que uma mudança desse tipo provavelmente alteraria totalmente o aminoácido codificado (Li & Graur, 1991). Os modelos propostos por Bonney utilizam apenas uma seqüência de DNA na análise. Assim sendo, o modelo se caracteriza de maneira simples, pois utiliza apenas a seqüência de um indivíduo na análise.

Pesquisas futuras poderiam tentar solucionar esses problemas, como por exemplo, desenvolver um modelo mais extenso, que tenha suposições mais plausíveis. Poderia-se propor um modelo em que a dependência se encontra nos nucleotídeos da terceira posição do códon e nos nucleotídeos da primeira posição do códon seguinte e utilizar mais do que uma seqüência de DNA, ou seja, vários indivíduos. Con-

siderando esse modelo, a primeira posição de um códon não depende da segunda e a segunda não depende da terceira, mas a terceira posição depende da próxima posição na seqüência, ou seja, o primeiro nucleotídeo do próximo códon.

Hipoteticamente, consideremos uma seqüência de DNA onde todas as posições são supostas independentes. Assim, teríamos apenas três logitos e faríamos o produtório para as k posições da seqüência. Esse modelo possuiria apenas três parâmetros a serem estimados, porém, sabemos que de todos os modelos possíveis para modelar dados de seqüência de DNA, esse seria o modelo menos adequado pois ele não se adequa às realidades biológicas. O outro extremo seria considerar que todas as posições da seqüência são dependentes, independentemente do tipo de dependência. Assim, o modelo teria $3k$ logitos, pois para cada posição da seqüência têm-se três logitos. É fácil notar que esse modelo é muito complexo pois possui um número de parâmetros gigantesco. Além disso, seria complicado propor uma estrutura de dependência para esse modelo. Para o caso do modelo que queremos desenvolver, temos $k + 6$ logitos (seis para a primeira e segunda posição do códon e $\frac{3k}{3}$ para a terceira posição), ou seja, ainda não é possível implementar esse modelo pois o número de logitos, e consequentemente, o número de parâmetros é muito grande.

Referências

Agresti, A. *Categorical Data Analysis*. Wiley Series in Probability and Mathematical Statistics, 1990. Applied Probability and Statistics, Hardcover.

Anderson, S., Bankier, A., and Barrell, B. Sequence and organization of the human mitochondrial genome. *Nature 290* (1981), 457–465.

Andrade, M., and Pinheiro, H. Métodos estatísticos aplicados em genética humana. *15 SINAPE, ABE* (2002).

Andrews, R., Kubacka, I., Chinnery, P., Lightowlers, R., Turnbull, D., and Howell, N. Reanalysis and revision of the cambridge reference sequence for human mitochondrial dna. *Nature Genet. 23* (1999), 147.

Bagchi, P., Jiang, O., and Bonney, G. Compound regressive models for quantitative multivariate phenotypes:application to lipid and lipoprotein data. *Genetic Epidemiology 10* (1993), 647–651.

Bahadur, R. A representation of the joint distribution of responses to n dichotomous items. *In studies in Item Analysis and Prediction* (1961), 158–176.

Bickel, P., and Doksum, K. *Mathematical Statistics, Basic Ideas and Selected Topics*. 2001.

Bonney, G. On the statistical determination of major gene mechanisms in continous human traits:regressive models. *American Journal of Medical Genetics 18* (1984), 731–749.

Bonney, G. Regressive logistic models for familial disease and other binary traits. *Biometrics 42* (1986), 611–625.

Bonney, G. Logistic regression for dependent binary observations. *Biometrics 43* (1987), 951–973.

Bonney, G. Compound regressive models for family data. *Hum Hered. 42* (1992), 28–41.

Bonney, G., Amfoh, K., and Shaw, R. The use of logistic models for the analysis of codon frequencies of dna sequences in terms of explanatory variables. *Biometrics 50* (1994), 1054–1063.

Bonney, G., Lathrop, G., and Lalouel, J. Combined linkage and segregation analysis using regressive models. *Am. J. Hum Genetic 43* (1988), 29–37.

Brooks, C., and Bonney, G. A simulation study of properties of a regressive logistic model. *Journal of Statistical Computation and Simulation 32* (1989), 31–43.

Casley, D. *Primer on Molecular Biology.* Technical report. Department of Energy, Officce of Health and Environment Research, U.S., 1992.

Cox, D. The analysis of multivariate binary data. *Applied Statistics 21* (1972), 113–120.

Elston, R. Segregation analysis. *In current Developments in Anthropological Genetics 1* (1980), 327–354. Edited by JH Mielke and MH Crawford.

Granthan, R. Amino acid difference formula to help explain protein evolution. *Science 185* (1974), 862–864.

Hosmer, D., and Lemeshow, S. *Applied Logistic Regression.* John Wiley and Sons, Inc., New York, 1989.

Lesaffre, E. *Logistic discriminant analysis with applications in electrocardiography.* PhD thesis, University of Leuven, Belgium, 1986.

Lewis, R. *Human Genetics-Concepts and Applications*, 4 ed. Mc Graw Hill, 2001.

Li, W.-H., and Graur, D. *Fundamentals of Molecular Evolution*, 1st ed. Sinauer Associates, Inc., Sunderland - Massachusetts, 1991.

Manistis, T., Fritsch, E., and Sambrook, J. *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbor Laboratory, Cold Springer Harbor, N.Y., 1982.

Orssaud, C. *Leber's hereditary optic neuropathy.* Orphanet Encyclopedia, novembro 2003.

Ott, J. *Analysis of Human Genetic Linkage.* The Johns Hopkins University Press, Baltimore and London, 1999.

Paula, G. Modelos de regressão com apoio computacional. Notas de aula, USP, 2004.

Rosner, B. Multivariate methods in ophthalmology with applications to other paired-data situations. *Biometrics 40* (1984), 1025–1035.

Stiratelli, R., Laird, N., and Ware, J. Random-effects models for serial observations with binary response. *Biometrics 40* (1984), 961–971.

Tavaré, S., and Giddings, B. Some statistical aspects of the primary structure of nucleotide sequences. *In Matematical Methods for DNA sequences, M.S. Waterman (ed).CRC Press* (1988), 117–132.

Apêndice

Programa para calcular a log-verossimilhança do modelo aditivo com covariáveis.

```
function pr = covariaveis(n)
B1=[0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 0 0 1 1 1 1 2 2 2 2 3 3 3
3 0 0 0 0 1 1 1 1 2 2 3 3 3 3];
B2=[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2
2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3];

B3=[0 1 2 3 0 1 2 3 0 1 2 3 0 1 2 3 0 1 2 3 0 1 2 3 0 1 2 3 0 1 0 1 2 3 0 1 2 3 0
1 2 3 0 1 2 3 0 1 2 3 0 1 2 3 0 1 0 1 2 3];

ts=[0 0 1 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 1 0 0 0 0 2 2 0 0 1 1 0 0 2 2 0
0 1 1 0 0 2 2 0 0 1 1 2 2 0 0 1 1];

aa=[57.67 57.67 33 33 39.78 39.78 40 40 50.56 50.56 38 38 48.44 48.44 42.11 42.11
71.33 71.33 61.44 61.44 46.33 46.33 46.22 46.22 40.67 40.67 34.78 34.78 48.22 48.22 46.11
46.11 82.89 82.89 53.89 53.89 40.22 40.22 75.78 75.78 52.22 52.22 80.67 80.67 55.67 55.67
154.67 154.67 105.81 105.81 72.11 72.11 63.22 63.22 58.22 58.22 67 67 64 64];

av=[9.6 9.6 9.1 9.1 9.1 9.1 9.1 9.1 8.8 8.8 8.5 8.5 8.4 8.4 8.4 8.4 10 10 10 10 8.6 8.6 8.6
8.6 7.3 7.3 7.3 7.3 9.4 9.4 9.4 9.4 8.9 8.9 7.5 7.5 7.7 7 9.7 9.7 9 9 11 11 9.1 9.1 16.8 16.8
11.6 11.6 8.5 8.5 8.5 8.5 10 10 10.4 10.4 10.4 10.4];

z11=[0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0];

z12=[0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 1 1 1 1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0];

z13=[0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0
1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1];
```

```
z21=[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0];

z22=[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0];

z23=[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1];

for i=1:60
if B1(i)==0
m11(i)=0; m12(i)=0; m13(i)=0; m14(i)=1;
elseif B1(i)==1
m11(i)=1; m12(i)=0; m13(i)=0; m14(i)=0;
elseif B1(i)==2
m11(i)=0; m12(i)=1; m13(i)=0; m14(i)=0;
elseif B1(i)==3
m11(i)=0; m12(i)=0; m13(i)=1; m14(i)=0;
end;
if B2(i)==0
m21(i)=0; m22(i)=0; m23(i)=0; m24(i)=1;
elseif B2(i)==1
m21(i)=1; m22(i)=0; m23(i)=0; m24(i)=0;
elseif B2(i)==2
m21(i)=0; m22(i)=1; m23(i)=0; m24(i)=0;
elseif B2(i)==3
m21(i)=0; m22(i)=0; m23(i)=1; m24(i)=0;
end;
if B3(i)==0
m31(i)=0; m32(i)=0; m33(i)=0; m34(i)=1;
elseif B3(i)==1
m31(i)=1; m32(i)=0; m33(i)=0; m34(i)=0;
elseif B3(i)==2
m31(i)=0; m32(i)=1; m33(i)=0; m34(i)=0;
```

```
elseif B3(i)==3
m31(i)=0; m32(i)=0; m33(i)=1; m34(i)=0;
end;
end;
a(1) = sym('a11','real');
a(2) = sym('a12','real');
a(3) = sym('a13','real');
a(4) = sym('a21','real');
a(5) = sym('a22','real');
a(6) = sym('a23','real');
a(7) = sym('a31','real');
a(8) = sym('a32','real');
a(9) = sym('a33','real');
b(1) = sym('b1','real');
b(2) = sym('b2','real');
b(3) = sym('b3','real');
s(1) = sym('s1','real');
s(2) = sym('s2','real');
s(3) = sym('s3','real');
s(4) = sym('s4','real');
s(5) = sym('s5','real');
s(6) = sym('s6','real');
s(7) = sym('s7','real');
s(8) = sym('s8','real');
s(9) = sym('s9','real');


    denomina=0;
for i= 1:60
t11=(a(1)+b(1)*aa(i)+b(2)*ts(i)+b(3)*av(i));
t12=(a(2)+b(1)*aa(i)+b(2)*ts(i)+b(3)*av(i));
t13=(a(3)+b(1)*aa(i)+b(2)*ts(i)+b(3)*av(i));
t21=(a(4)+s(1)*z11(i)+s(2)*z12(i)+s(3)*z13(i)+b(1)*aa(i)+b(2)*ts(i)+b(3)*av(i));
```

```
t22=(a(5)+s(1)*z11(i)+s(2)*z12(i)+s(3)*z13(i)+b(1)*aa(i)+b(2)*ts(i)+b(3)*av(i));
t23=(a(6)+s(1)*z11(i)+s(2)*z12(i)+s(3)*z13(i)+b(1)*aa(i)+b(2)*ts(i)+b(3)*av(i));
t31=(a(7)+s(4)*z11(i)+s(5)*z12(i)+s(6)*z13(i)+s(7)*z21(i)+s(8)*z22(i)+
s(9)*z23(i)+b(1)*aa(i)+b(2)*ts(i)+b(3)*av(i));
t32=(a(8)+s(4)*z11(i)+s(5)*z12(i)+s(6)*z13(i)+s(7)*z21(i)+s(8)*z22(i)+
s(9)*z23(i)+b(1)*aa(i)+b(2)*ts(i)+b(3)*av(i));
t33=(a(9)+s(4)*z11(i)+s(5)*z12(i)+s(6)*z13(i)+s(7)*z21(i)+s(8)*z22(i)+
s(9)*z23(i)+b(1)*aa(i)+b(2)*ts(i)+b(3)*av(i));
lpcodon(i)= t11*m11(i)+t12*m12(i)+ t13*m13(i)-
log(1+ exp(t11)+exp(t12)+exp(t13))+t21*m21(i)+t22*m22(i)+ t23*m23(i)-
log(1+ exp(t21)+exp(t22)+exp(t23))+ t31*m31(i)+t32*m32(i)+ t33*m33(i)-
log(1+ exp(t31)+exp(t32)+exp(t33));
end;
lvero= dot(n,lpcodon);
lvero
```

Tabela A.15: Número e Posição de mutações encontradas comparando a SRC com indivíduos doentes e não-doentes.

Observação	Indivíduo	Condição	Número de Diferenças	Posições
1	ay063349	leber	2	960;1.019
2	ay063350	leber	3	576;706;1.019
3	ay063351	leber	5	105;114;576;960;1.019
4	ay063352	leber	3	576;960;1.019
5	ay063353	leber	2	960;1.019
6	ay063354	leber	6	114;576;900;960;1.003;1.019
7	ay063355	leber	4	576;960;1.019;1.155
8	ay063356	normal	3	576;706;960
9	ay063357	leber	4	540;576;960;1.019
10	ay063358	leber	4	114;576;960;1.019
11	ay063359	normal	2	576;960
12	ay063361	normal	0	
13	ay063362	normal	1	960
14	ay063363	normal	1	960
15	ay063364	normal	3	114;723;960
16	ay063365	normal	2	576;960
17	ap008270	normal	3	114;576;960
18	ap008416	normal	3	114;576;960
19	ap008471	normal	5	42;576;777;888;960
20	ap008593	normal	3	114;576;960

Tabela A.16: Continuação.

Observação	Indivíduo	Condição	Número de Diferenças	Posições
21	ap008679	normal	2	576;960
22	ap008746	normal	3	114;576;960
23	ap008276	normal	5	42;576;777;888;960
24	ap008282	normal	3	114;576;960
25	ap008290	normal	5	42;576;777;888;960
26	ap008294	normal	3	114;576;960
27	ap008301	normal	3	114;576;960
28	ap008302	normal	3	114;576;960
29	ap008303	normal	3	114;576;960
30	ap008304	normal	3	114;576;960
31	ap008307	normal	2	576;960
32	ap008309	normal	4	114;456;576;960
33	ap008310	normal	3	114;576;960
34	ap008319	normal	3	114;576;960
35	ap008320	normal	3	114;576;960
36	ap008322	normal	4	576;777;888;960
37	ap008324	normal	5	42;576;777;888;960
38	ap008326	normal	3	114;576;960
39	ap008337	normal	3	114;576;960
40	ap008339	normal	3	114;576;960
41	ap008340	normal	4	576;777;888;960
42	ap008341	normal	3	114;576;960
43	ap008343	normal	5	42;576;777;888;960
44	ap008348	normal	3	114;576;960
45	ap008349	normal	5	42;576;777;888;960
46	ap008350	normal	3	114;576;960
47	ap008352	normal	4	114;456;576;960
48	ap008356	normal	5	114;456;576;771;960
49	ap008358	normal	3	114;576;960
50	ap008360	normal	4	114;576;960;1.155
51	ap008363	normal	5	42;576;777;888;960
52	ap008376	normal	7	114;256;325;408;576;942;960

Tabela A.17: Frequências Observadas da SRC do gene NADH4 e Valores das Covariáveis

Códons	Freq.	TSCORE	AARISK	AVDIST	Códons	Freq.	TSCORE	AARISK	AVDIST
TTT	9	0	57,67	9,6	TAT	2	2	82,89	8,9
TTC	11	0	57,67	9,6	TAC	11	2	82,89	8,9
TTA	8	1	33,00	9,1	TAA	0	*	*	*
TTG	1	1	33,00	9,1	TAG	0	*	*	*
CTT	10	0	39,78	9,1	CAT	1	0	53,89	7,5
CTC	31	0	39,78	9,1	CAC	12	0	53,89	7,5
CTA	42	0	40,00	9,1	CAA	9	1	40,22	7,7
CTG	4	0	40,00	9,1	CAG	1	1	40,22	7,0
ATT	16	0	50,56	8,8	AAT	2	0	75,78	9,7
ATC	23	0	50,56	8,8	AAC	21	0	75,78	9,7
ATA	24	1	38,00	8,5	AAA	10	2	52,22	9,0
ATG	3	1	38,00	8,5	AAG	1	2	52,22	9,0
GTT	0	0	48,44	8,4	GAT	0	0	80,67	11,0
GTC	4	0	48,44	8,4	GAC	3	0	80,67	11,0
GTA	8	0	42,11	8,4	GAA	9	1	55,67	9,1
GTG	1	0	42,11	8,4	GAG	0	1	55,67	9,1
TCT	5	0	71,33	10,0	TGT	1	0	154,67	16,8
TCC	17	0	71,33	10,0	TGC	2	0	154,67	16,8
TCA	10	1	61,44	10,0	TGA	12	2	105,81	11,6
TCG	1	1	61,44	10,0	TGG	1	2	105,81	11,6
CCT	3	0	46,33	8,6	CGT	0	0	72,11	8,5
CCC	14	0	46,33	8,6	CGC	5	0	72,11	8,5
CCA	6	0	46,22	8,6	CGA	4	1	63,22	8,5
CCG	0	0	46,22	8,6	CGG	0	1	63,22	8,5
ACT	8	0	40,67	7,3	AGT	2	2	58,22	10,0
ACC	17	0	40,67	7,3	AGC	8	2	58,22	10,0
ACA	22	1	34,78	7,3	AGA	0	*	*	*
ACG	1	1	34,78	7,3	AGG	0	*	*	*
GCT	6	0	48,22	9,4	GGT	1	0	67,00	10,4
GCC	12	0	48,22	9,4	GGC	9	0	67,00	10,4
GCA	8	0	46,11	9,4	GGA	4	1	64,00	10,4
GCG	0	0	46,11	9,4	GGG	3	1	64,00	10,4

Tabela A.18: Probabilidades Observadas e Estimadas para os modelos com Covariáveis para n= 30 seqüências

Probabilidade	Modelos			
Observada	Independente	Igual. Pred.	Markov	Aditivo
0,0196	0,0203	0,0183	0,0165	0,0176
0,0240	0,0504	0,0345	0,0324	0,0346
0,0173	0,0537	0,0348	0,0313	0,0338
0,0022	0,0049	0,0032	0,0028	0,0031
0,0219	0,0311	0,0413	0,0437	0,0267
0,0674	0,0623	0,0824	0,0822	0,0912
0,0918	0,0549	0,0726	0,0724	0,0803
0,0086	0,0050	0,0066	0,0066	0,0073
0,0349	0,0191	0,0303	0,0236	0,0463
0,0501	0,0651	0,0736	0,0752	0,0667
0,0523	0,0565	0,0638	0,0652	0,0581
0,0065	0,0051	0,0058	0,0059	0,0053
0	0,0063	0,0007	0,0017	0,0009
0,0087	0,0263	0,0075	0,0071	0,0068
0,0174	0,0246	0,0076	0,0069	0,0069
0,0022	0,0022	0,0007	0,0006	0,0006
0,0109	0,0085	0,0137	0,0194	0,0202
0,0370	0,0245	0,0332	0,0320	0,0317
0,0218	0,0218	0,0299	0,0287	0,0283
0,0022	0,0020	0,0027	0,0026	0,0026
0,0054	0,0117	0,0068	0,0103	0,0061
0,0316	0,0400	0,0306	0,0288	0,0315
0,0131	0,0351	0,0269	0,0253	0,0276
0	0,0032	0,0024	0,0023	0,0025
0,0174	0,0097	0,0088	0,0098	0,0196
0,0371	0,0663	0,0698	0,0698	0,0671
0,0480	0,0597	0,0621	0,0611	0,0596
0,0021	0,0054	0,0056	0,0055	0,0054
0,0131	0,006	0,0049	0,0136	0,0098
0,0261	0,0127	0,0242	0,0204	0,0217
0,0174	0,0104	0,0205	0,0173	0,0183
0	0,0009	0,0019	0,0016	0,0017
0,0045	0,0005	0,0008	0,0005	0,0002
0,0238	0,0070	0,0071	0,0096	0,0072
0,0022	0,0047	0,0039	0,0017	0,0006
0,0266	0,0419	0,0390	0,0387	0,0422

Tabela A.19: Continuação

Probabilidade	Modelos			
Observada	Independente	Igual. Pred.	Markov	Aditivo
0,0196	0,0325	0,0286	0,0287	0,0317
0,0022	0,0035	0,0034	0,0033	0,0036
0,0027	0,0078	0,0118	0,0039	0,0060
0,0474	0,0315	0,0227	0,0256	0,0244
0,0219	0,0316	0,0254	0,0269	0,0275
0,0022	0,0029	0,0023	0,0024	0,0025
0	0,0038	0,0049	0,0044	0,0022
0,0065	0,0074	0,0111	0,0109	0,0114
0,0195	0,0121	0,0179	0,0177	0,0189
7,26216E-05	0,0011	0,0016	0,0016	0,0017
0,0022	0,0101	0,0087	0,0064	0,0049
0,0044	0,0040	0,0015	0,0022	0,0013
0,0261	0,0075	0,0124	0,0150	0,0144
0,0022	0,0007	0,0011	0,0014	0,0013
0	0,0032	0,0010	0,0005	0,0002
0,0105	0,0253	0,0224	0,0211	0,0231
0,0087	0,0220	0,0191	0,0177	0,0197
0	0,0020	0,0017	0,0016	0,0018
0,0044	0,0053	0,0029	0,0011	0,0020
0,0174	0,0164	0,0121	0,0120	0,0114
0,0022	0,0024	0,0011	0,0011	0,0007
0,0196	0,0046	0,0085	0,0085	0,0084
0,0107	0,0047	0,0081	0,0078	0,0079
0,0046	0,0004	0,0007	0,0007	0,0007

Tabela A.20: Probabilidades Observadas e Estimadas para os modelos sem Covariáveis para n= 30 seqüências

Probabilidade	Modelos			
Observada	Independente	Igual. Pred.	Markov	Aditivo
0,0196	0,0116	0,0148	0,0108	0,0153
0,0240	0,0359	0,0221	0,0237	0,0216
0,0173	0,0345	0,0215	0,0234	0,0212
0,0022	0,0031	0,0019	0,0021	0,0019
0,0219	0,0166	0,0300	0,0325	0,0218
0,0674	0,0514	0,0734	0,0715	0,0768
0,0918	0,0494	0,0716	0,0703	0,0754
0,0086	0,0045	0,0065	0,0064	0,0068
0,0349	0,0195	0,0223	0,0246	0,0312
0,0501	0,0606	0,0559	0,0542	0,0511
0,0523	0,0583	0,0545	0,0533	0,0502
0,0065	0,0053	0,0049	0,0048	0,0045
0,0000	0,0079	0,0034	0,0049	0,0046
0,0087	0,0246	0,0115	0,0107	0,0108
0,0174	0,0236	0,0112	0,0105	0,0106
0,0022	0,0021	0,0010	0,0010	0,0010
0,0109	0,0077	0,0113	0,0115	0,0145
0,0370	0,0239	0,0278	0,0280	0,0265
0,0218	0,0230	0,0270	0,0276	0,0261
0,0022	0,0021	0,0025	0,0025	0,0024
0,0054	0,0110	0,0056	0,0086	0,0049
0,0316	0,0343	0,0226	0,0208	0,0226
0,0131	0,0329	0,0221	0,0205	0,0222
0,0000	0,0030	0,0020	0,0019	0,0020
0,0174	0,0130	0,0103	0,0160	0,0179
0,0371	0,0404	0,0424	0,0390	0,0381
0,0480	0,0388	0,0413	0,0383	0,0374
0,0021	0,0035	0,0037	0,0035	0,0034
0,0131	0,0053	0,0042	0,0085	0,0070
0,0261	0,0164	0,0233	0,0208	0,0216
0,0174	0,0158	0,0227	0,0205	0,0212
0,0000	0,0014	0,0021	0,0019	0,0019
0,0045	0,0054	0,0079	0,0023	0,0030
0,0238	0,0167	0,0198	0,0232	0,0228
0,0000	0,0160	0,0193	0,0228	0,0224
0,0000	0,0015	0,0017	0,0021	0,0020
0,0022	0,0077	0,0039	0,0017	0,0009
0,0266	0,0239	0,0161	0,0172	0,0176
0,0196	0,0229	0,0157	0,0170	0,0173
0,0022	0,0021	0,0014	0,0015	0,0016
0,0027	0,0091	0,0071	0,0032	0,0037

Tabela A.21: Continuação

Probabilidade	Modelos			
Observada	Independente	Igual. Pred.	Markov	Aditivo
0,0474	0,0281	0,0302	0,0323	0,0320
0,0219	0,0270	0,0294	0,0318	0,0315
0,0022	0,0025	0,0027	0,0029	0,0029
0,0000	0,0037	0,0029	0,0017	0,0014
0,0065	0,0114	0,0166	0,0172	0,0174
0,0195	0,0110	0,0162	0,0170	0,0171
0,0001	0,0010	0,0015	0,0015	0,0015
0,0022	0,0036	0,0041	0,0021	0,0028
0,0044	0,0113	0,0140	0,0152	0,0148
0,0261	0,0109	0,0136	0,0149	0,0146
0,0022	0,0010	0,0012	0,0014	0,0013
0,0000	0,0052	0,0020	0,0016	0,0009
0,0105	0,0162	0,0112	0,0113	0,0116
0,0087	0,0155	0,0109	0,0111	0,0114
0,0000	0,0014	0,0010	0,0010	0,0010
0,0044	0,0061	0,0037	0,0030	0,0034
0,0174	0,0191	0,0210	0,0211	0,0209
0,0000	0,0183	0,0205	0,0207	0,0205
0,0000	0,0017	0,0019	0,0019	0,0019
0,0022	0,0025	0,0015	0,0016	0,0013
0,0196	0,0077	0,0115	0,0113	0,0114
0,0107	0,0074	0,0112	0,0111	0,0112
0,0046	0,0007	0,0010	0,0010	0,0010